The OSCAR PUBLISHING Services
American Journal of Applied Science and Technology

VOLUME Vol.06 Issue03 2026
PAGE NO. 118-124
DOI 10.37547/ajast/Volume06Issue03-23

Algorithm for Assessing the Quality of Medical Synthetic Data Based on The Multi-Criteria and Pac-Bayesian Model

Juraev Gulomjon Primovich

Associate Professor of the Department of "Exact Sciences, Land Cadastre and Municipal Services" of the International Innovation University, Doctor of Technical Sciences (PhD), Uzbekistan

Jovlieva Dilnoz Mustofa kizi

Lecturer of the Department of "Exact Sciences, Land Cadastre and Utilities" of the International Innovation University, Uzbekistan

Received: 12 February 2026; **Accepted:** 08 March 2026; **Published:** 31 March 2026

Abstract: The effectiveness of medical artificial intelligence systems is directly related to the quality and level of representativeness of the training sample. However, real clinical data are characterized by problems such as confidentiality limitations, class mismatch, and heterogeneity. In solving these problems, synthetic data is considered a promising solution, however, a comprehensive assessment of their quality remains a pressing issue.

In this work, an algorithm for assessing the quality of synthetic data based on a multi-criteria approach and PAC-Bayesian theory is proposed. The proposed model combines statistical proximity (KL-divergence), inter-character relationships (mutual information), predictive efficiency, and subject-specific limitations within a single integrated functional.

Experimental results showed that the synthetic data satisfactorily reflect the main features of the real distribution (KL = 0.6035), and inter-characteristic relationships are maintained with high accuracy (MI = 0.0072). The model, trained on the basis of synthetic data, achieved an accuracy of 80.4%, which confirms the possibility of its practical application. The domain matching criterion showed the maximum value (1.0).

The results show that the proposed approach ensures a balance between the quality of synthetic data and model reliability.

Keywords: Synthetic data, medical artificial intelligence, multi-criteria assessment, PAC-Bayes, KL-divergence, mutual information, predictive efficiency, bootstrap generation.

Introduction: In recent years, the rapid development of artificial intelligence (AI) and machine learning technologies has brought the processes of diagnosis, prediction, and individualized treatment in the healthcare system to a qualitatively new level [11-12]. Approaches based on medical data are becoming an important factor in supporting clinical decision-making, early detection of diseases, and the formation of personalized treatment strategies for patients [13]. However, the effectiveness and reliability of these systems are directly determined by the quality, completeness, and level of representativeness of educational information [11].

Working with real clinical data is associated with a number of fundamental limitations. In particular, since medical data includes sensitive personal information, they require strict confidentiality requirements [6-7]. These limitations limit the possibility of large-scale use of real data and complicate the processes of developing and testing machine learning models.

In addition, real clinical data sets are often characterized by a heterogeneous structure, which is small in size, disproportionately distributed by classes, and includes continuous, categorical, and ordered features [13]. These factors reduce the generalizability of models and create problems of reliability in their practical application.

Synthetic data generation is widely used as a promising approach to solving these problems [14-15]. In particular, the Generative Adversarial Networks (GAN) model demonstrates effective results in the study of real-world data distributions [1], the Wasserstein GAN (WGAN) allows for the stabilization of this process [2], the Variational Autoencoder (VAE) offers a probabilistic generative mechanism [3], while the CTGAN was developed as an adapted model for tabular data [4]. These approaches allow for the generation of synthetic datasets while preserving the statistical properties of real data.

At the same time, the problem of assessing the quality of synthetic data has not yet been fully solved. In existing studies, assessment is often carried out on the basis of individual criteria - statistical proximity (KL-divergence) [7], predictive effectiveness [11], or the risk of confidentiality (for example, membership inference attacks) [8]. However, these criteria are interconnected, and their separate application does not provide a complete and objective picture of the quality of synthetic data [10, 15]. In particular, increasing confidentiality often reduces the usefulness of information, which creates a fundamental balance problem between confidentiality and efficiency [14].

Moreover, the relationship between the quality of synthetic data and the generalizability of machine learning models is insufficiently theoretically substantiated. In solving this problem, the PAC-Bayesian approach, which is one of the important directions of statistical learning theory, is of great importance [11]. This approach allows us to determine the upper limits for the generalization error of the model, but it has not yet been sufficiently applied in the context of synthetic data.

In this study, the generation of synthetic data was carried out not through complex generative models, but based on the bootstrap (re-sampling) approach. Although this approach allows for the preservation of the statistical properties of real data, it does not generate new information and, as a result, can introduce limitations on predictive effectiveness. Therefore, the quality of synthetic data should be assessed comprehensively not only on the basis of statistical criteria, but also from the point of view of their practical usefulness.

Taking into account the above-mentioned problems, this article proposes an algorithm for assessing the quality of medical synthetic data based on the multi-criteria and PAC-Bayesian model. The proposed algorithm combines statistical proximity, inter-characteristic relationships, predictive efficiency, and limitations inherent in the subject area within a single integrated evaluation mechanism. Also, based on the PAC-Bayesian theory, the relationship between the generalization reliability of the model and the quality of synthetic data is theoretically substantiated.

METHOD

This section describes the approach to solving the above-mentioned pressing issues. In particular, the following aspects are considered: problem statement, multi-criteria integrated model, criterion of statistical proximity, criterion of maintaining inter-characteristic relationships, criterion of predictive effectiveness, as well as limitations and logical consistency inherent in the subject area. Also, an analysis is carried out based on the PAC-Bayesian model, and finally, the optimization problem is solved.

Based on these models and criteria, an algorithm for assessing the quality of medical synthetic data has been developed. Based on the developed algorithm, experimental results were obtained and analyzed.

1. Formulation of the issue

Suppose a real data set is defined as follows:

$$D = \{x_1, x_2, \dots, x_n\}, x_i \sim P(X)$$

and synthetic data set:

$$\widehat{D} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_m\}, \hat{x}_i \sim Q(X)$$

Here:

- $P(X)$- real data distribution;
- $Q(X)$- synthetic distribution studied by the generative model.

Medical information is usually presented in a high-dimensional, mixed-type (continuous, categorical, and

ordered) feature space. Therefore, the quality of synthetic data should be assessed taking into account not only statistical proximity, but also inter-character relationships and domain restrictions.

Problem 1. It is required to assess the quality of the distribution $Q(X)$ based on the following criteria:

- $D_{KL}(P \parallel Q) \to min$
- $I_P(X_i, X_j) \approx I_Q(X_i, X_j)$
- $Acc(_{f_{\hat{D}}}, D_{test}) \to max$
- $C(x) = 1$

2. Multi-criteria integrated model

The quality of synthetic data for solving a pressing issue identified in the literature review is determined by the following multi-criteria functional:

$$Q_{total}(P, Q) = w_1 Q_{stat} + w_2 Q_{dep} + w_3 Q_{pred} + w_4 Q_{logic} \quad (1)$$

Here:

$Q_{stat} = D_{KL}(P \parallel Q)$ – criterion of statistical proximity;

$Q_{dep} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |I_P(X_i, X_j) - I_Q(X_i, X_j)|$ criterion of inter-character dependence;

$Q_{pred} = Acc(_{f_{\hat{D}}}, D_{test})$ –predictive effectiveness;

$Q_{logic} = \mathbb{E}_{x \sim Q}[C(x)]$ –compliance with domain restrictions.

w_i –are weight coefficients that determine the equilibrium between the criteria [13-14] and the weight coefficients satisfy the following conditions:

$$\sum_{i=1}^{4} w_i = 1, w_i \geq 0$$

This model, unlike existing approaches, integrates all criteria into a single system.

3. Criterion of Statistical Proximity

The difference between real and synthetic distributions is estimated using Kullback-Leibler divergence:

$$Q_{stat} = D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)} \quad (2)$$

This criterion serves to bring the global statistical properties of synthetic data as close as possible to real data. The smaller the value of the KL-divergence, the closer the distributions are considered [12].

4. Criterion for maintaining interrelationships

In medical data, the semantic and statistical relationship between signs is of great importance. This relationship is evaluated through mutual information:

$$Q_{dep} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |I_P(X_i, X_j) - I_Q(X_i, X_j)| \quad (3)$$

where mutual information is defined as follows:

$$I(X_i, X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} P(x_i, x_j) log \frac{P(x_i, x_j)}{P(x_i) P(x_j)}$$

Here:

- I_P –mutual information for real data;
- I_Q –Mutual information for synthetic data.

This criterion represents the degree of preservation of structural relationships between features and ensures that the internal structure of synthetic data corresponds to real data.

5. Criterion of predictive effectiveness

The practical value of synthetic data is determined by the effectiveness of the model trained on their basis in real data. Therefore, the following criterion is introduced:

$$Q_{pred} = Acc(_{f_{\hat{\theta}}}, D_{test}) \quad (4)$$

here:

- $f_{\hat{\theta}} = \mathcal{A}(D)$ –model trained on the basis of synthetic data;
- $D_{test} \approx P(x)$ –a test set consisting of real data;
- $Acc(\cdot)$ –is the classification accuracy[16], which is defined as follows:

$$Acc(f, D_{test}) = \frac{1}{|D_{test}|} \sum_{(x,y) \in D_{test}} 1(f(x) = y)$$

This criterion assesses the ability of a model trained on the basis of synthetic data to generalize in real data.

6. Subject-specific limitations and logical consistency

Medical data have specific semantic and logical limitations, which are determined on the basis of clinical knowledge. If these limitations are violated, the synthetic data will not correspond to the real medical context, and the reliability of the model will sharply decrease. Therefore, an important component in

assessing the quality of synthetic data is the consideration of subject-specific limitations.

In this work, the restrictions specific to the subject area are formalized through the following indicator function:

$$Q_{logic} = \mathbb{E}_{x\sim Q}[C(x)]C(x) = \begin{cases} 1, & x \in \Omega \\ 0, & aks\ holda \end{cases}, \quad (5)$$

Here:

Ω –space of medically permissible values;

$C(x)$ –An indicator function that checks logical consistency.

This component ensures the clinical consistency of synthetic data.

7. Analysis based on the PAC-Bayer model

In this work, the effectiveness of the machine learning model **trained based on synthetic data** is assessed based on the PAC-Bayesian theory:

$$R(\rho) \le \hat{R}(\rho) + \sqrt{\frac{D_{KL}(\rho\|\pi)+\ln(1/\delta)}{2N}} \quad (6)$$

Here:

- $R(\rho)$ –true error;
- $\hat{R}(\rho)$ –empirical error;
- π –prior distribution;
- ρ –posterior distribution;
- $D_{KL}(\rho \parallel \pi)$ –KL-divergence;
- N –number of training samples;
- $\delta \in (0,1)$ –reliability parameter;

This expression allows us to theoretically substantiate the relationship between the quality of synthetic data and the reliability of the model in real conditions [15].

8. Final optimization problem

As a result, the quality of synthetic data is determined by the following optimization problem:

$$Q^* = \arg\min_Q Q_{total}(P, Q) \quad (7)$$

Here Q_{pred} and Q_{logic}, since the and criteria must have a maximum value, they are taken into account in the process of optimization with a negative sign.

This boundary estimate makes it possible to theoretically substantiate the reliability of the model obtained on the basis of an optimized distribution Q^*.

In this study, the generation of synthetic data was carried out on the basis of the bootstrap method. That is, new samples were created by random selection from a real dataset. This approach allows maintaining the statistical properties of the real distribution at a high level.

However, since the bootstrap method does not generate new information, it imposes certain limitations on the predictive effectiveness of synthetic data. As a result, a balance is observed between statistical proximity and practical utility.

9. Proposed algorithm

The algorithm for assessing the quality of medical synthetic data based on the multi-criteria and PAC-Bayesian model consists of the following stages:

1. Based on real data, synthetic data is generated using the bootstrap method;
2. The difference between the real and synthetic data distributions is estimated using Kullback-Leibler divergence;
3. Mutual information values are determined for character pairs;
4. Based on synthetic data, a machine learning model is trained, and its accuracy is assessed using real test data;
5. Compliance with the restrictions of the subject area is checked;
6. All criteria are combined into a single assessment based on a weighted sum;
7. Based on the PAC-Bayesian theory, the reliability of the model's operation is assessed;
8. A final quality assessment is formed.

This algorithm allows for a comprehensive assessment of the quality of synthetic data.

Then, based on the steps of this algorithm described above, the sequence of steps of the algorithm is written as follows:

Step 1. Synthetic data is generated using the bootstrap method based on real data;

Step 2. The difference between the real and synthetic

distributions is determined by formula Q_{stat} (2);

Step 3. The mutual information difference for character pairs is calculated using formula Q_{dep} (3);

Step 4. Based on synthetic data, the model $f_{\widehat{D}}$ is trained and evaluated according to formula Q_{pred} (4);

Step 5. The process of checking compliance with the subject area is determined based on formula Q_{logic} (5)**;**

Step 6. When forming a multi-criteria functional, all criteria are combined into a single functional, i.e., $Q_{total}(P, Q)$ is calculated based on formula (1);

Here, each criterion reflects an important aspect of the quality of synthetic data, and their combination provides a comprehensive character of the overall assessment.

Step 7. The problem of optimizing the synthetic distribution is determined by the formula Q^* (7) given above;

Step 8. The reliability of the model is assessed based on the PAC-Beiß criterion, i.e., based on formula $R(\rho)$ (6);

Step 9. Based on the quality of synthetic data and model reliability, the final assessment is determined and the process is stopped.

As a result, the proposed algorithm allows for a comprehensive assessment of the quality of synthetic data based on statistical proximity, inter-character relationships, predictive efficiency, and domain constraints, and also theoretically substantiates the reliability of the model's generalization.

RESULT AND DISCUSSION

1. Data set

In this study, the UCI Heart Disease Dataset [16] for the diagnosis of cardiovascular diseases was used. This dataset is one of the standard reference data sets widely used for medical diagnostics.

The dataset includes 14 key attributes, including: age, sex, type of chest pain (cp), blood pressure (trestbps), cholesterol levels (old man), heart rate (thalach), and other clinical indicators. The target indicates the presence of a variable heart disease.

In the study, the dataset was pre-processed, categorical characters were converted to digital format, and the target variable was converted to binary form.

2. Experimental setup

Experimental studies were conducted in the following stages:

- The dataset was divided into 85% training and 15% test sets;
- The symbols were brought to the same scale through standardization (normalization);
- Synthetic data were generated using the bootstrap (re-sampling) method;
- The Random Forest classification model was trained based on a synthetic dataset;
- The model was evaluated based on real test data;
- As evaluation criteria, KL-divergence, mutual information difference, accuracy, domain consistency, and PAC-Bayesian boundary were used.

In order to ensure the reliability of the experimental results, all experiments were repeated 100 times. In each iteration, the data were randomly divided into 85% training and 15% test sets. The final results were assessed based on average values. This approach allows us to assess the stability of the model results and determine the influence of randomness.

3. Experimental results

The obtained results are presented in Table 1 below:

Table 1

Experimental results

Criterion	Value
KL divergence	0.6035
Mutual information difference	0.0072
Accuracy	80.4%
Domain match	1.0

PAC Bayes	0.6862
Q_total	-0.2983

The experimental results show the effectiveness of the proposed model. The value of the KL-divergence (0.6035) indicates that the difference between the synthetic and real distributions is moderate. At the same time, the difference in inter-character relationships is very small (0.0072), which indicates that synthetic data successfully preserved the internal structural features of real data.

According to the predictive effectiveness criterion, the model showed a high result (80.4%). This confirms that the model trained on the basis of synthetic data also works effectively in real data. The maximum value of the domain correspondence criterion indicates the complete correspondence of the synthetic data to the subject area.

Also, based on experiments repeated 100 times, the standard deviation of accuracy was 0.0539, which indicates the stability of the model results.

DISCUSSION

The results of the proposed model show that the use of unified criteria for assessing the quality of synthetic data is insufficient. As observed in the experimental results, although the statistical proximity (KL-divergence) was at an average level, the interrelationships (mutual information) were preserved with high accuracy. This indicates that the internal structure of synthetic data is close to real data.

A significantly higher predictive effectiveness (80.4%) confirms the possibility of using synthetic data in practical tasks. At the same time, the maximum value of the domain compatibility criterion indicates the clinical validity of the data generated by the model.

As can be seen from the results, although the bootstrap-based approach is effective in storing statistical properties, the possibility of generating new information is limited. Therefore, there is a certain balance between statistical proximity and predictive effectiveness.

The PAC-Bayesian boundary indicates that the model's generalizability is at a satisfactory level. In general, the proposed multi-criteria approach allows for a comprehensive assessment of the quality of synthetic data.

CONCLUSION

In this study, the problem of assessing the quality of medical synthetic data was studied comprehensively based on a multi-criteria and PAC-Bayesian approach. The proposed model made it possible to assess statistical proximity, inter-character relationships, predictive effectiveness, and limitations inherent in the subject area using a single integrated functional.

The experimental results confirmed the effectiveness of the model. In particular, the high accuracy of maintaining inter-character relationships (MI = 0.0072) and the achievement of a high level of predictive effectiveness (80.4%) indicate the practical value of synthetic data. The maximum value of the domain compatibility criterion confirms the clinical correctness of the generated data.

At the same time, the average level of statistical convergence and the PAC-Bayesian boundary indicate a balance between the quality of synthetic data and the generalizability of the model.

The results expand the possibilities of applying the proposed approach in medical artificial intelligence systems and can serve as an effective tool for assessing the quality of synthetic data.

REFERENCES

1. I.Goodfellow, J.Pouget-Abadie, M.Mirza, B.Xu, D.Warde-Farley, S.Ozair, A.Courville and Y.Bengio. "Generative Adversarial Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672-2680.
2. M.Arjovsky, S.Chintala and L.Bottou. "Wasserstein GAN," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 214-223.
3. D.P.Kingma and M.Welling. "Auto-Encoding Variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

4. L.Hu, M.Skoularidou, A.Cuesta-Infante and K.Veeramachaneni. "Modeling Tabular Data using Conditional GAN," in *Advances in Neural Information Processing Systems (NeurIPS)* , 2019.
5. J.Yoon, J.Jordon and M.van der Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees," in *International Conference on Learning Representations (ICLR)* , 2019.
6. C. Dwork "Differential Privacy" in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, 2006, pp. 1-12.
7. C.Dwork and A.Roth. *The Algorithmic Foundations of Differential Privacy.* Boston, MA, USA: Now Publishers.
8. R.Shokri, M.Stronati, C.Song and V.Shmatikov. "Membership Inference Attacks Against Machine Learning Models," in *IEEE Symposium on Security and Privacy*, 2017, pp. 3-18.
9. A.Esteban, S.L.Hyland and G.Rätsch. "Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs," in *Machine Learning for Healthcare Conference,* 2017, pp. 276-296.
10. A.Borji. "Pros and Cons of GAN Evaluation Measures," *Computer Vision and Image Understanding,* vol. 179, pp. 41-65, 2019.
11. E. J.Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Natural Medicine,* vol. 25, pp. 44-56, 2019.
12. R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics,* vol. 19, no. 6, pp. 1236-1246, 2018.
13. R.Chen, J.Lu and Z.Chen, "Synthetic Data in Healthcare: A Survey," *Journal of the American Medical Informatics Association,* vol. 28, no. 11, pp. 2497-2508, 2021.
14. B.Jayaraman and D.Evans. "Evaluating Differentially Private Machine Learning in Practice," in *USENIX Security* Symposium, 2019, pp. 1895-1912.
15. A.El Emam, "Seven ways to evaluate the utility of synthetic data," *IEEE Security & Privacy,* vol. 18, no. 3, pp. 56-62, 2020.
16. UCI Machine Learning Repository, "Heart Disease Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease.