Methods for Calculating Reliability, Representativeness, and Confidentiality Scores for Artificial Data Generation

I) Reliability

The reliability of the generated data will be assessed by a panel of medical experts. The experts will review the artificial data to determine its compliance with clinical trends observed in real data. Specific criteria include the consistency of diagnoses, treatments, and reported patient outcomes.

Definition and Thresholds:

- Very Good Reliability: Synthetic data is almost indistinguishable from real data, with clinical consistency above 95%.
- Average Reliability: Synthetic data shows minor but acceptable discrepancies, with clinical consistency between 75% and 95%.
- Poor Reliability: Synthetic data shows significant discrepancies, with clinical consistency below 75%.

II) Representativeness

To evaluate representativeness, several statistical measures will be used:

1. Mean (μ) of each variable of the patient vector:
   $\mu = (1/N) \Sigma x_i$

   where x_i represents the data values and N is the total number of samples.

2. Standard Deviation (σ) for all variables independently:
   $\sigma = \sqrt{((1/N) \Sigma (x_i - \mu)^2)}$

   where x_i represents the data values and μ is its mean.

3. Skewness:
   $\text{Skewness} = (1/N) \Sigma [(x_i - \mu) / \sigma]^3$

   where x_i represents the data values, μ is its mean, and σ is its standard deviation.

4. Kurtosis:
   Kurtosis = (1/N) Σ [(x_i - μ) / σ]^4 - 3

   where x_i represents the data values, μ is its mean, and σ is its standard deviation.

5. Frobenius distance between Covariance Matrices

   The Frobenius distance between two covariance matrices measures the difference between the covariance structures of two multivariate Gaussian distributions. It is particularly useful for comparing the spread and interrelationships between variables in real and synthetic datasets.

   The Frobenius distance between two covariance matrices Σ_{real} and $\Sigma_{\text{synthetic}}$ is defined as:

$$\text{Frobenius Distance} = \|\Sigma_{\text{real}} - \Sigma_{\text{synthetic}}\|_F$$

   where $\|\cdot\|_F$ denotes the Frobenius norm, which is calculated as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}|^2}$$

   for a matrix $A = [a_{ij}]$ of size $n \times m$.

   The Frobenius distance between covariance matrices is a critical metric for assessing how well the synthetic data maintains the multivariate relationships and variability present in the real data. A low Frobenius distance indicates that the synthetic data closely mirrors the covariance structure of the real data, ensuring that the synthetic data retains the dependencies and variability essential for accurate analysis and modeling. Conversely, a high Frobenius distance suggests significant differences in the covariance structures, potentially compromising the synthetic data's utility for reliable inference and decision-making.

   Conditional Means and covariances:
   Comparison of conditional means and covariances on relevant data subsets. The tolerance is 5%.

Acceptance Thresholds:

- First and second-order moments (mean and standard deviation): maximum deviation of 5%
- Analysis of the differences for third-order and fourth-order moment (hard to determine a good threshold, it is strongly case dependent)
- Covariance matrices Frobenius distance: below 1
- Conditional means: tolerance of 5%

In addition to moment measures, three tests will be applied

- Kolmogorov-Smirnov Test (KS Test):
- Wasserstein Distance
- Jensen-Shannon Divergence (JS Divergence)

The three tests (Kolmogorov-Smirnov, Wasserstein Distance, and Jensen-Shannon Divergence) are complementary because they evaluate different characteristics of data distributions, providing a more comprehensive assessment of the representativeness of synthetic data:

1. Kolmogorov-Smirnov Test (KS Test)

   The KS test compares the cumulative distribution of synthetic data $F_n(x)$ to that of real data $F(x)$:

   $D = \sup_x |F_n(x) - F(x)|$

   where sup represents the supremum.

   The KS test compares the cumulative distributions of synthetic and real data. It is particularly sensitive to differences in the shapes of the cumulative distribution functions. The KS test detects discrepancies in the overall shape of the distributions, which is crucial for identifying global differences between synthetic and real data.

2. Wasserstein Distance

   $W(p, q) = \inf_{\gamma \in \Pi(p,q)} E_{(x,y)\sim\gamma}[|x - y|]$

   where p and q are the distributions of real and synthetic data respectively, and $\Pi(p, q)$ is the set of couplings γ of p and q.

   The Wasserstein distance measures the minimum "cost" to transform one distribution into another, considering the entire distributions. Unlike the KS test, the Wasserstein distance accounts for the distance between each point in the

distributions, offering a more intuitive and holistic measure of differences between the distributions.

3. Jensen-Shannon Divergence (JS Divergence)

$$\text{JS}(P||Q) = \frac{1}{2}[D_{\text{KL}}(P||M) + D_{\text{KL}}(Q||M)]$$
$$D_{\text{KL}}(P||Q) = \sum P(i) \log \left[\frac{P(i)}{Q(i)}\right]$$

where $M = \frac{1}{2}(P + Q)$ and D_{KL} is the Kullback-Leibler divergence.

The JS divergence evaluates the similarity between two probability distributions. It is symmetric and finite, making it a stable and interpretable measure. The JS divergence measures shared information between the distributions and identifies differences in terms of probability, complementing the structural measures provided by the KS test and the Wasserstein distance.

By using these three tests together, a complete view of the differences between synthetic and real data is obtained, covering aspects of overall shape, global distance, and probabilistic similarity.

Acceptance Thresholds:

- **Very Good Representativeness**: Tolerance below 5% for all metrics and statistical tests.
- **Good Representativeness**: Tolerance between 5% and 10% for all metrics and statistical tests.
- **Poor Representativeness**: Tolerance above 10% for all metrics and statistical tests

III) Confidentiality

To evaluate confidentiality, the following distances will be used:

1. Euclidean Distance

$$d_E(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

- $d_E(x, y)$: Euclidean distance between points x and y
- x_i and y_i: Components of vectors x and y in the original and synthetic data, respectively

The Euclidean distance measures the straight-line distance between two points in a multidimensional space. This metric provides a straightforward and intuitive measure of overall differences between original and synthetic data points, capturing the geometric distance in the feature space. A higher Euclidean distance indicates that synthetic data points are sufficiently distant from the original data points, ensuring that individual records are not easily re-identifiable, thus maintaining confidentiality.

2. Manhattan Distance:

$$d_M(x, y) = \sum |x_i - y_i|$$

- $d_M(x, y)$: Manhattan distance between points x and y
- x_i and y_i: Components of vectors x and y in the original and synthetic data, respectively

The Manhattan distance measures the distance between two points by summing the absolute differences of their coordinates. This metric is less sensitive to outliers compared to Euclidean distance and provides a robust measure of differences by considering each dimension independently. A higher Manhattan distance implies that there are significant differences in each individual dimension, ensuring that synthetic data points are distinct from the original data points in multiple aspects, thereby protecting confidentiality.

3. Cosine Distance:

$$d_C(x, y) = 1 - \frac{x \cdot y}{||x||\,||y||}$$

- $d_C(x, y)$: Cosine distance between points x and y
- $x \cdot y$: Dot product of vectors x and y
- $||x||$ and $||y||$: Euclidean norms of vectors x and y

The cosine distance measures the cosine of the angle between two vectors, focusing on their orientation rather than magnitude. This metric highlights the similarity in direction between the original and synthetic data points, making it effective for understanding how well the relationships between features are preserved. A higher cosine distance indicates that the synthetic data points have different orientations compared to the original data points, ensuring that the relative feature relationships in the synthetic data are not directly traceable to the original data, thus maintaining confidentiality.

4. Nearest Neighbor Distance Ratio (NNDR):

$$\text{NNDR} = \frac{d_{\text{nearest}}(x_{\text{original}}, x_{\text{synthetic}})}{d_{\text{nearest}}(x_{\text{original}}, x_{\text{original}})}$$

where d_{nearest} represents the nearest neighbor distance.

The NNDR is used to verify if the synthetic data is sufficiently different from the original data. This involves comparing the distance between each original data point and its nearest synthetic neighbor with the distance between the same original point and its nearest original neighbor. The NNDR is calculated as follows:

By using these four metrics together, we can comprehensively demonstrate that synthetic data points are sufficiently distant from the original data points, ensuring that confidentiality is maintained. This combination of metrics covers absolute geometric distance, dimension-specific differences, and directional similarity, providing a robust assessment of the protection of individual data records.

Acceptable Thresholds:

- **Very Good Confidentiality**: above 1 for Euclidean and Manhattan distances, above 0.3 for Cosine distance, and NNDR tolerance below 0.01.
- **Good Confidentiality**: between 0.5 and 1for Euclidean and Manhattan distances, between 0.2 and 0.3 for Cosine distance, and NNDR tolerance between 0.01 and 0.05.
- **Poor Confidentiality**: below 0.5 for Euclidean and Manhattan distances, below 0.2 for Cosine distance, and NNDR tolerance above 0.05.

Construction of Synthetic Patient Samples

Our method uses deep neural networks to generate high-quality synthetic health data. We employ a model called Variational Autoencoder (VAE), which imposes a probability distribution on the latent space, allowing for more flexible and controlled data sampling.

Details of the Method Used

1. Variational Autoencoders (VAE):

   The VAE improves upon traditional autoencoders by incorporating generative hierarchical statistical models, specifically Mixed Effect Models, where observations are driven by latent conditional probabilities. Neural networks add flexibility beyond classical Mixed Effect Models, as the latent variables are designed to follow a parametric probability distribution. This feature allows for more flexible and controlled data sampling.

2. Latent Space Geometry:

   Our VAE model treats the latent space as a Riemannian manifold, enhancing the reliability, precision, and robustness of sampling. The inherent structure of a Riemannian manifold facilitates the definition of distances, geodesics (shortest paths), and interpolation. This approach improves the diversity and accuracy of the generated data by capturing the geometrical and probabilistic structure of the data in a lower-dimensional but representative space.

3. Riemannian Sampling:

   We employ Hamiltonian Monte Carlo (HMC) sampling, which leverages the Riemannian geometry of the latent space for more efficient exploration. This method captures the subtle nuances of health data, enhancing the quality and representativeness of the synthetic data. By accounting for the curvature and structure of the Riemannian manifold, the HMC algorithm provides more accurate and diverse samples, ensuring that the synthetic data closely mirrors the complexity of real-world health data.