Supplementary materials: The Scorecard for Synthetic Medical Data Evaluation and Reporting

"When a measure becomes a target, it ceases to be a good measure."

– Goodhart's Law

This paper introduces a comprehensive, multidimensional framework for evaluating synthetic data. Inspired by Goodhart's Law, it underscores the risks of over-reliance on one metric and instead advocates for holistic evaluation across congruence, coverage, constraints, completeness, consistency, compliance, and comprehension—the 7Cs. By systematically measuring these dimensions, the paper enables stakeholders to assess the quality of synthetic data and its readiness for the intended application.

A. Approaches for Evaluating Synthetic Medical Data

We categorize approaches for evaluating synthetic medical data (SMD) into three main types: intrinsic data quality evaluation (task-independent), task-dependent evaluation, and human-dependent evaluation (Table 1).

Intrinsic data quality evaluation focuses on assessing the intrinsic quality of the synthetic dataset itself, independent of any specific downstream task. This involves evaluating the data in the image, raw data, or feature space based on criteria such as congruence, coverage, and constraint adherence, as outlined in the evaluation framework. This approach allows for an early indication of potential quality issues that could later affect task-specific performance. For instance, if general data quality evaluation reveals low fidelity or inadequate coverage, this suggests that the dataset may lack critical patterns or diversity, which might result in poor generalization and subpar performance in downstream applications. The task-independent approach is advantageous for identifying data quality issues early, but it requires robust feature representation and quantification to be effective.

Task-dependent evaluation measures the quality of synthetic data based on its effectiveness for a specific task. This method typically involves training a model on synthetic data and then testing it on real data, or vice versa, to assess whether the synthetic data leads to better training outcomes or improved testing results. This approach provides detailed insights into the utility of the synthetic data and helps determine its effectiveness in supporting the intended purpose. However, its generalizability is limited because it evaluates the synthetic data's quality solely within the context of that specific task. Furthermore, this approach tends to identify issues late in the evaluation process, which can delay necessary improvements. Such late detection is costly as it requires revisiting and retraining models to enhance the datasets. This process can become cyclical if the root causes of poor task performance are not fully understood. Consequently, task-independent evaluation, which involves a deeper and comprehensive assessment of the data based on well-defined criteria such as correctness, coverage, and constraints, is needed for ensuring broader data quality.

Human-dependent visual inspection is a subjective evaluation method, where domain experts visually compare synthetic and real data to assess quality. This approach is valuable for initial quality checks and can capture subtle, qualitative aspects that automated metrics might overlook.

Table 1: Summary of Evaluation Approaches for SMD

Approach	Scope	Task-Specific?	Subjective?	Identify Early Issues?	Scalable?	Quantitative?	Feature/Data Representation
Intrinsic Quality (7Cs)	Dataset / Sample	×	×	✓	✓	✓	Required for quantitative metrics (e.g., FID)
Task-Specific	Dataset	✓	×	×	✓	✓	Inherent in model architecture
Human-Dependent	Sample	×	✓	×	×	×	Not required

However, it is inherently subjective, inconsistent, and not scalable, as it relies on individual interpretation and may miss minor statistical deviations.

Each evaluation approach—intrinsic SMD quality, task-specific, and human-dependent—has unique strengths and limitations. For optimal assessment, these approaches should be combined rather than used in isolation. Starting with intrinsic data quality evaluation provides a broad overview and can uncover early issues that may affect downstream tasks. Following this with task-specific evaluations allows us to measure SMD utility, and check if early quality issues, such as inadequate coverage, align with task-specific performance limitations. Human-dependent visual inspection can further validate SMD. This combination ensures a comprehensive evaluation, where synthetic datasets are systematically refined through quantitative metrics, task performance, and human expertise.

B. Intrinsic Data Quality Evaluation (Criteria & Metrics)

Table 2 provides an organized summary of metrics for evaluating synthetic data quality, grouped according to key evaluation criteria—Congruence, Coverage, Constraint, Completeness, Compliance, and Consistency—which collectively assess the suitability of synthetic data for its intended purpose. The "Point/Set" column specifies whether each metric operates at the level of individual data points or entire sets; the "Unary/Binary" column indicates whether the metric assesses the synthetic data alone (unary) or in comparison to real data (binary); the "Space" column describes the operational space for each metric, such as embedding,raw data, meta data, or image/sample space; the "Direction" column identifies the optimal direction for each metric (e.g., maximize or minimize) to indicate whether higher or lower values represent better performance.

B.1 Congruence

The Congruence category evaluates how well synthetic data approximates the real data distribution. Key metrics **jordon2022synthetic** summarized in the table include Cosine Similarity (CS), which measures vector alignment between real and synthetic distributions, with values close to 1 indicating high similarity in the feature embeddings. Earth Mover's Distance (EMD) captures the distance between distributions, where lower values signify closer alignment, while Jensen-Shannon Divergence (JSD) quantifies distribution similarity, with values near 0 indicating high congruence. Peak Signal-to-Noise Ratio (PSNR) assesses fidelity in image or pixel-level data, where higher values reflect close resemblance to real images, and Structural Similarity

Index Measure (SSIM) evaluates image structure, with values close to 1 indicating preserved characteristics. Fréchet Inception Distance (FID) measures statistical similarity in an embedding space, with lower scores implying a closer match to real data. Finally, Precision (Prec) assesses how accurately synthetic data captures real features, with higher values suggesting a high-fidelity match.

B.2 Coverage

The Coverage (Novelty & Diversity) metrics evaluate the extent to which synthetic data spans the feature space, and can be used to assess the diversity across the dataset and the inclusion of novel yet valid instances. Examples of Coverage metrics are summarized in Table 2. One of the mostly used metric is the Inception Score (IS), which measures diversity by evaluating how well synthetic samples match real-world data in terms of variety and quality **salimans2016improved**. It works by passing synthetic samples through a pre-trained Inception model, which generates probability distributions over possible classes for each sample. The IS measures both the diversity of these samples (how varied the generated samples are across classes) and their quality (how confidently the model classifies them into specific categories). Recall (Rec) quantifies the extent to which real data samples are represented within the synthetic dataset by constructing a nearest-neighbor manifold based on synthetic data and determining how many real data samples fall within this synthetic data manifold. Higher Recall values indicate better representation of real data within the synthetic data distribution **alaa2022faithful**. In contrast, Coverage (Cov) assesses how well the synthetic dataset captures the diversity of real data by constructing nearest-neighbor manifolds around real samples, ensuring that each neighborhood in the real data includes at least one synthetic sample. Unlike Recall, which examines real data in terms of synthetic data coverage, Coverage directly compares synthetic data with real data by focusing on real data neighborhoods **alaa2022faithful**.

The Distance to Centroid (DtC) metric **regenwetter2023beyond** can be used to assess both diversity and novelty within a synthetic dataset by calculating how far each data point is from a central point, such as the centroid or geometric median. Points that are moderately distant from the centroid reflect variety within the dataset, suggesting that it captures a range of valid features without being overly concentrated around a single type. However, it's essential to recognize that very far data points might indicate outliers rather than genuine diversity or novelty. Excessive distance from the centroid could suggest errors, noise, or unrealistic samples that deviate from the core distribution of the dataset. This potential for outliers underscores the need for careful interpretation of distance values, ideally using this metric in conjunction with others to differentiate between true diversity and irrelevant samples.

Another metric for measuring diversity is the Convex Hull Volume (CHV), which measures the volume around synthetic samples tp capture the dataset's spread in feature space. Higher CHV values indicate broader coverage, but the metric is sensitive to outliers, which can inflate the volume **sharafutdinov2022application**. Determinantal Point Processes (DPPs) score has been used recently to evaluate diversity of synthetic data **regenwetter2023beyond** by examining the eigenvalues of a matrix based on distances between samples within the generated dataset. The determinant of this matrix represents the volume spanned by the points in feature space, effectively capturing both the spread and uniformity of data points. However, this score is sensitive to duplicates—if any eigenvalues are zero, the determinant collapses, indicating zero

diversity. Additionally, the nonlinearity of the DPP diversity score means small changes in score can reflect significant shifts in diversity, making it highly sensitive to variations in sample structure.

Vendi Score (VS) is a reference-free metric that captures the effective number of unique patterns or modes in the dataset by calculating entropy-based dissimilarity **dan2023vendi**. Variance (Var) measures the spread of feature values within the synthetic dataset, and captures the range and variability of its data points. Jaccard Index (JI) evaluates the overlap between categorical features in real and synthetic data, which is particularly useful for ensuring the synthetic dataset includes a representative range of categories or labels, such as demographic groups or disease types. Entropy measures the disorder or unpredictability of features within the dataset, where high entropy reflects a balanced distribution of features or classes, preventing over-representation of any single subset and maintaining diversity. Finally, Rarity Score (RS) is a novelty metric that highlights unique synthetic samples using k-Nearest Neighbors (k-NN) to assess how distant each synthetic sample is from the real data distribution. A higher RS implies samples lie in lower-density regions of the feature space, indicating novelty without frequent overlap with real data **han2022rarity**. Unlike congruence metrics, which have clear optimization directions (e.g., minimize or maximize), coverage metrics depend on the specific task, as excessively high values may suggest unrealistic diversity, reducing dataset reliability. We note that while all coverage metrics have a general direction of Maximize, the ideal level of maximization should be adjusted based on the specific task requirements. Excessively high values may lead to overly diverse or unrealistic synthetic data, so each metric's maximization must align with the intended application.

B.3 Constraint

To evaluate constraint adherence in SMD, several metrics outlined in Table 2 can be applied. One approach involves assessing the proximity of synthetic data points to known invalid regions. For example, the Nearest Invalid Datapoint (NID) **regenwetter2023beyond** metric calculates the distance of each sample to the nearest constraint-violating data point from a reference set of invalid examples. Lower NID values suggest that samples are closer to these constraint-violating points, indicating a higher likelihood of boundary violations, while higher values imply a safer distance within valid regions. Following NID, Signed Distance to Constraint Boundary (SDCB) **regenwetter2023beyond** offers another way to quantify constraint adherence. SDCB calculates the distance from each sample to a mathematically defined constraint boundary, with the unique feature of a "signed" distance. Positive values indicate samples are within valid regions, while negative values highlight constraint violations. This metric provides insight into how strongly each sample aligns with or diverges from predefined constraints, which can be particularly useful when a continuous boundary is available for constraints.

Another metric, the Constraint Satisfaction Rate (CSR), measures the proportion of samples in the synthetic dataset that meet all specified constraints, while its complement, Constraint Violation Rate (CVR), captures the percentage of samples that fail to satisfy at least one constraint. In medical datasets, CSR might represent criteria such as anatomical validity (e.g., each breast having one nipple) or disease validity (e.g., symptoms aligning with clinical guidelines).

Additionally, classifier-based metrics can be used to identify constraint violations within synthetic datasets. Here, a classifier trained on valid and invalid samples according to established

boundaries or a labeled valid dataset generates metrics such as Validity Precision (CV-P), Validity Recall (CV-R), and False Positive Rate for Constraint Violations (CV-FPR). These metrics assess the classifier's accuracy in detecting samples that meet or violate known constraints.

We note that the majority of constraint metrics depend on predefined constraint boundaries or standards, often derived from anatomical structures, clinical guidelines, or application-specific standards. However, if a reference dataset of invalid data points/samples is available, it can be used to refine or verify these boundaries for more accurate constraint adherence assessment.

B.4 Completeness

To assess completeness in SMD, several metrics can be adopted from data quality to evaluate whether all necessary components for specific medical applications are included. For example, one metric could be the Proportion of Required Metadata Fields (PRMF), which calculates the percentage of required metadata fields present in the dataset. If an application requires ten specific metadata fields and only eight are included, the score would be 80%. Another metric could be the Missing Data Percentage (MDP), which calculates the percentage of missing values across all fields by dividing the number of missing values by the total number of values in the dataset. Similarly, another metric could be Labeled Data Percentage (LDP) to measure the percentage of data points that are labeled. For textual data, the Missing Sections in Text (MST) could be used as a metric to evaluate the average number of missing sections, particularly relevant for structured reports or notes. In the context of LLMs, completeness refers to the extent to which a model's output comprehensively addresses a given prompt or query and covers all relevant aspects, necessary details, and critical information.

B.5 Compliance

To evaluate compliance in SMD, several metrics can be used to assess alignment with privacy and standardization requirements, as summarized in Table 2. For example, the Differential Privacy Score (DPS) **ee** quantifies the level of noise required to meet differential privacy standards. A lower DPS indicates that less noise is needed to protect individual privacy, implying that the synthetic data is already less susceptible to re-identification risks. Another metric to evaluate privacy is the K-Anonymity Level (KA) **ee**, which measures the degree to which an individual's information is indistinguishable from that of k-1 other individuals in the dataset. Higher KA values indicate stronger privacy protection, as each record is masked among more similar records, making it harder to identify individuals. For instance, a KA level of 5 means that each record in the dataset has at least four other records with identical attributes, making it difficult to single out an individual. The L-Diversity Score (LD) **ee** enhances K-Anonymity by ensuring that sensitive information within a dataset is diverse, even among groups with similar attributes. Higher LD scores indicate that each group of indistinguishable records (from KA) contains varied values for sensitive attributes, protecting against homogeneity attacks. For example, if all individuals within a group share the same disease status, they are still identifiable despite KA protections. A high LD score means that sensitive attributes vary within groups. The T-Closeness Level (TC) **ee** further strengthens privacy by ensuring that the distribution of sensitive attributes in any group of records closely mirrors the overall dataset distribution. This prevents attackers from gaining insight into sensitive information even when groups are

formed by similar attributes. A high TC score indicates that each group's sensitive attributes are proportionally distributed, reducing the chance of inferring sensitive details. While there are currently no widely established metrics specifically for measuring synthetic data formatting and standardization, some practical approaches can be drawn from data quality and interoperability standards. For example, metrics can be adopted from data consistency measures, which focus on assessing consistency in data formatting—such as standardized date formats, numerical ranges, and categorical encodings **Batini2006**; **HIMSS2020**.

B.6 Consistency

Consistency can be measured by evaluating the stability and uniformity of data quality across subgroups and over time using a combination of statistical and variability-based metrics. Average variability metrics, such as variance, capture the overall stability of quality metrics such as congruence, coverage, constraint satisfaction, or completeness across groups (e.g., age, ethnicity, disease class), with lower variance indicating greater consistency. Extreme variation metrics, such as the max-min difference, identify worst-case inconsistencies by measuring the disparity between the best and worst-performing groups, such as coverage rates across subgroups. Statistical tests, such as ANOVA, can be employed to detect significant differences in quality metrics between groups, indicating potential inconsistencies if subgroup means differ substantially. For example, a low p-value in an ANOVA test for correctness scores across demographic groups would suggest that the quality of SMD is not uniform across those groups. Longitudinal consistency can be assessed by tracking changes in data quality metrics over time to ensure stability as data landscapes evolve.

B.7 Comprehension

The clarity of documentation and interpretable rules can be used for assessing the comprehension of SMD generation methods. Clarity of documentation evaluates how transparently and comprehensively the data generation process is described, including the underlying assumptions, algorithms, validation, and workflows. High-quality documentation ensures that different users, such as clinicians and researchers, can understand the methodology, identify potential limitations, and trust the generated data for medical applications. For instance, generative adversarial networks (GANs) typically exhibit lower comprehension due to their black-box nature, where the underlying logic of the synthetic data creation process is not easily interpretable. In contrast, knowledge-based methods often have higher comprehension because they rely on explicit, rule-based models that are inherently explainable, allowing users to understand how specific characteristics, such as tumor size or shape, are synthesized. To assess and score comprehension for each method, a combination of metrics can be used, such as a Documentation Clarity Score (DCS) to measure the transparency and completeness of the documentation, and a Rule Interpretability Score (RIS) to evaluate the extent to which interpretable rules can describe the data generation process. GANs would score lower on both metrics due to their opaque processes, while knowledge-based methods would score higher for their well-defined rules and explicit workflows.

Category	Metric	Point/Set	Unary/Binary	Space	Direction
Congruence (Fidelity)	CS	Set	Binary	Embedding	Maximize
	EMD	Set	Binary	Embedding	Minimize
	JSD	Set	Binary	Embedding	Minimize
	PSNR	Point	Binary	Image/Sample	Maximize
	SSIM	Point	Binary	Image/Sample	Maximize
	IS	Set	Unary	Embedding	Maximize
	FID	Set	Binary	Embedding	Minimize
	Prec	Set	Binary	Embedding	Maximize
Coverage (Novelty & Diversity)	IS		Binary	Embedding	Maximize
	Rec	Set	Binary	Embedding	Maximize
	Cov	Set	Unary	Embedding	Maximize
	DtC	Point	Unary	Embedding	Maximize
	CHV	Set	Unary	Embedding	Maximize
	VS	Set	Unary	Embedding	Maximize
	Var	Set	Unary	Embedding	Maximize
	JI	Set	Binary	Embedding	Maximize
	Ent	Set	Unary	Embedding	Maximize
	RS	Point	Unary	Embedding	Maximize
Constraint	NIP	Point	Unary	Image/Sample	Minimize
	CSR	Set	Unary	Image/Sample	Maximize
	CV-P	Set	Unary	Embedding	Maximize
	CV-R	Set	Unary	Embedding	Maximize
	CV-DR	Set	Unary	Embedding	Maximize
	CV-FPR	Set	Unary	Embedding	Minimize
Completeness	PRMF	Set	Binary	Metadata Fields	Maximize
	MDP	Set	Unary	Feature space	Minimize
	LDP	Set	Unary	Label Space	Maximize
	MST	Point	Unary	Textual Data	Minimize
Compliance	DPS	Set	Unary	Data Attribute	Minimize
	KA	Point	Unary	Data Attribute	Maximize
	LD	Point	Unary	Data Attribute	Maximize
	TC	Point	Unary	Data Attribute	Maximize
Consistency	Variance	Set	Unary	Quality Metrics	Minimize
	Max-Min Dif.	Set	Unary	Quality Metrics	Minimize
	ANOVA	Set	Unary	Quality Metrics	Stat. Significance
Comprehension	DCS	–	Unary	Documentation	Maximize
	RIS	–	Unary	Model	Maximize

Table 2: Table 2 summarizes some metrics to evaluate synthetic data quality across key dimensions: Congruence, Coverage, Constraint, Completeness, Compliance, Consistency, and Comprehension. **Category** groups metrics based on evaluation criteria (e.g., Congruence or Completeness); **Metric** lists the specific metric, with abbreviations aligned to definitions provided in the text and summarized in Table 3; **Point/Set** specifies whether the metric operates at the level of individual data points (Point) or aggregates over the entire dataset (Set); **Unary/Binary** indicates whether the metric assesses the synthetic data alone (Unary) or compares it to real data (Binary); **Space** describes the operational space for each metric, such as Embedding (feature space representations), Metadata Fields (e.g., patient demographics), Model Space (e.g., generation method), or Documentation; and **Direction** identifies the optimization goal for the metric, where "Maximize" means higher scores are better, "Minimize" means lower scores are better, and "Stat. Significance" applies to metrics like ANOVA that detect meaningful differences.

Table 3: Abbreviations of Metrics for Synthetic Data Quality Evaluation

Abbreviation	**Full Name/Description**
CS	Cosine Similarity
EMD	Earth Mover's Distance
JSD	Jensen-Shannon Divergence
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
FID	Fréchet Inception Distance
Prec	Precision
IS	Inception Score
Rec	Recall
Cov	Coverage
DtC	Distance to Centroid
CHV	Convex Hull Volume
DPP	Determinantal Point Processes Score
VS	Vendi Score
Var	Variance
JI	Jaccard Index
Ent	Entropy
RS	Rarity Score
NID	Nearest Invalid Datapoint
SDCB	Signed Distance to Constraint Boundary
CSR	Constraint Satisfaction Rate
CVR	Constraint Violation Rate
CV-P	Validity Precision
CV-R	Validity Recall
CV-FPR	False Positive Rate for Constraint Violations
PRMF	Proportion of Required Metadata Fields
MDP	Missing Data Percentage
LDP	Labeled Data Percentage
MST	Missing Sections in Text
DPS	Differential Privacy Score
KA	K-Anonymity Level
LD	L-Diversity Score
TC	T-Closeness Level
DCS	Documentation Clarity Score
RIS	Rule Interpretability Score

C. Scorecard Example

The **Synthetic Data Scorecard Template** (Table 4) provides a structured framework for documenting and evaluating synthetic medical datasets. It organizes information into sections to ensure clarity and comprehensiveness for users. Below is an overview of the main sections of the scorecard:

- **General Information**: Captures essential details, including the dataset's name, release date, size, provenance, intended use, labels, and the method of generation (e.g., generative AI or knowledge-based). It also includes the dataset's impact on research, citation, licensing details, and contact information.
- **Data Quality (Quantitative Results)**: Evaluates the dataset's quality using the 7Cs framework: Congruence, Coverage, Constraint, Completeness, Compliance, Comprehension, and Consistency. Each dimension provides a quantitative measure of alignment between the synthetic dataset and its reference dataset.
- **Task-based Evaluation**: Assesses the dataset's performance on specific tasks, including relevant metrics such as accuracy, precision, recall, and F1-score.
- **Human-based Evaluation (Qualitative Results)**: Captures results from qualitative studies, such as reader evaluations, and observations about the dataset's realism, utility, and failure cases.
- **Synthetic Dataset Training and Validation Process**: Provides transparency about the generation, validation, and testing methods used to create the synthetic dataset.
- **Reference Dataset Information**: Documents details about the reference dataset (if applicable), including its size, clinical population, acquisition devices, and preprocessing methods.

Table 4: Synthetic Medical Data (SMD) Card Template

Synthetic Data General Information	
Name	[Dataset name]
Release Date	[Release date]
Dataset Size	[Total size of the dataset]
Dataset Provenance	[Origin and generation method of the dataset]
Dataset Intended Use	[Applications and use cases]
Dataset Labels	[Dataset labels]
Dataset Impact	[Impact on research and applications]
Dataset Citation	[Dataset citation]
Licensing Information	[License details]
Point of Contact	[Contact details for inquiries]
Synthetic Data Quality (Quantitative Results)	
Congruence	[Result and interpretation of congruence]
Coverage	[Result and interpretation of coverage]
Constraint	[Result and interpretation of constraint]
Completeness	[Result and interpretation of completeness]
Compliance	[Result and interpretation of compliance]
Comprehension	[Result and interpretation of comprehension]
Consistency	[Result and interpretation of consistency]
Task-based Evaluation (Quantitative Results)	
Task Performance	[Evaluation of dataset performance on specific tasks]
Task-specific Metrics	[Metrics such as accuracy, precision, recall, and F1-score for task evaluation]
Human-based Evaluation (Qualitative Results)	
Study Design	[Summary of the qualitative or reader study design]
Qualitative Evaluation	[Summary of qualitative results]
Reader Study Results	[Results of any reader study assessing dataset (e.g., utility, realism)]
Observations & Failure Cases	[Key insights from qualitative evaluations or reader studies]
Ethical Considerations, Limitations, and Recommendations	
Ethical Considerations	[Ethical concerns or considerations]
Biases	[Known biases or potential issues]
Limitations	[Known limitations of the dataset]
Recommendations	[Recommendations for dataset usage and best practices]
Synthetic Dataset Usage	
Directions for Use	• Repository: [Link to the repository] • Downloading: [Instructions for downloading the dataset] • Updating: [Instructions for updating the dataset] • Data Format: [Details of data format and preprocessing requirements] • Documentation: [Link to documentation]
Synthetic Dataset Training & Validation Process	
Generation Method	[Overall description of the training and generation process]
Validation Process	[Details of the validation process]
Testing Process	[Details of the testing process]
Reference Dataset General Information	
Purpose	[Purpose of the reference dataset, its intended use, and the problem it addresses]
Origin & Source	[Origin of the reference dataset, sites, and dates from which data was collected]
Dataset Size	[Number of images and number of patients in the reference dataset]
Clinical Population	[Description of the clinical population represented by the reference dataset]
Acquisition Devices	[Devices used for data acquisition in the reference dataset]
Reference Standard	[Details of the reference standard used for the dataset, including expert qualifications or annotation instructions]
Metadata	[Metadata available such as age, gender, breast density, etc.]
Ground Truth Labels	[Ground truth information or labels available for the dataset]
Preprocessing	[Preprocessing steps applied to the reference dataset]
Known Limitations	[Known limitations of the reference dataset]

D. Synthetic Data Card: Sinkove Dataset Example

D.1 Synthetic Data Card (Descriptive Section)

Table 5 and Table 6 summarized key information about Sinkove synthetic digital mammography dataset, which includes its intended use, labels, usage, and the dataset used for training and validation among others.

D.2 Synthetic Data Card (Quantitative and Qualitative Results)

The quality of Sinkove is summarized below across seven key criteria: Congruence D.2.1, Coverage D.2.2, Constraint D.2.3, Completeness D.2.4, Compliance D.2.5, Comprehension D.2.6, and Consistency D.2.7. As no human-based evaluations were conducted to assess Sinkove's quality, qualitative results are not included. However, Section **??** highlights examples of network-induced artifacts.

D.2.1 Congruence

Congruence measures how well the synthetic dataset aligns with the real dataset in terms of structural and visual similarity and feature representation. This dimension assesses whether the synthetic data adequately captures the patterns and characteristics of the real dataset.

Metrics and Justification
We used the following metrics to measure the alignment between Sinkove (synthetic) and CSAW-M (reference). Each metric was selected to capture a specific aspect of congruence:

- **Structural Similarity Index (SSIM, ↑)**
  Why SSIM? SSIM evaluates the preservation of spatial structures by comparing luminance, contrast, and structural similarity between images. It helps assessing whether the synthetic data captures the spatial characteristics of the reference dataset.
  Desirable Behavior: Higher SSIM indicates better structural similarity, meaning the synthetic dataset retains critical patterns from the real dataset.

- **Peak Signal-to-Noise Ratio (PSNR, ↑)**
  Why PSNR? PSNR quantifies the quality of image reconstruction by measuring the ratio of the maximum signal power to the noise power. It is used to evaluate how well the synthetic images approximate the reference dataset in terms of pixel-wise accuracy.
  Desirable Behavior: Higher PSNR reflects lower reconstruction error, implying higher fidelity to the original dataset.

- **Jensen-Shannon Divergence (JSD, ↓)**
  Why JSD? JSD measures the overlap between probability distributions. For this evaluation, it compares feature distributions extracted from the synthetic and real datasets, quantifying how well the synthetic dataset reproduces the statistical properties of the real dataset. In our case, the feature space was extracted using VGG16.
  Desirable Behavior: Lower JSD indicates better alignment of feature distributions.

Table 5: Synthetic Medical Data (SMD) Card - (Descriptive Section -Part 1)

Synthetic Data General Information	
Name	Sinkove Synthetic CSAW 100k Mammograms
Release Date	October 11, 2023
Dataset Size	100k synthetic mammograms
Dataset Provenance	Synthetic data generated using the CSAW-M subset and MONAI framework; i.e., Latent Diffusion Model (LDM) trained on the CSAW-M dataset.
Intended Use	Enhance AI model training for cancer masking in mammography
Labels	• Low Masking Level (score ≤ 2) • Medium Masking Level ($2 <$ score ≤ 6) • High Masking Level (score > 6)
Impact	Facilitates research and development of AI models for mammography
Citation	Pinaya, W. H., et al. (2023). Generative AI for medical imaging: extending the MONAI framework. *arXiv preprint arXiv:2307.15208*.
Licensing Information	Released under the Open & Responsible AI license (OpenRAIL).
Point of Contact	walter.diaz$_s$*anz@kcl.ac.uk*
Synthetic Data Quality (Quantitative Results)	
Congruence	Full results in Section D.2.1.
Coverage	Full Results in Section D.2.2.
Constraint	Full Results in Section D.2.3.
Completeness	Full Results in Section D.2.4.
Compliance	Full Results in Section D.2.4.
Consistency	Full Results in Section D.2.7.
Comprehension	Full Results in Section D.2.6.
Task-based Evaluation (Quantitative Results)	
Task Performance	Cancer masking classification (mixed Sinkove with CSAW for training).
Task-specific Metrics	Sensitivity: 0.84, Specificity: 0.83, F1-score: 0.85.
Human-based Evaluation (Qualitative Results)	
Study Design	Not Provided.
Qualitative Evaluation	Not Provided.
Observations & Failure Cases	See Figure **??**.
Ethical Considerations, Limitations, and Recommendations	
Ethical Considerations	Dataset adheres to HIPPA and privacy regulations; no personal identifiers present.
Biases	Underrepresentation of rare breast imaging features.
Recommendations	Combine with real datasets.
Usage	
Directions for Use	• **Repository:** The dataset can be accessed at Sinkove Synthetic CSAW Repository. • **Downloading:** Use the command `git clone https://huggingface.co/SinKove/mammography_csaw` to download the repository. Alternatively, the dataset can be downloaded directly as a compressed file from the provided link. • **Updating:** Run `git pull` within the cloned repository directory to update the dataset to the latest version. • **Data Format:** Ensure compatibility with AI frameworks by converting or preprocessing the data if necessary. Preprocessing scripts are available in the repository under the `scripts/` directory. • **Documentation:** Detailed usage instructions and examples can be found in the `README.md` file in the repository.

- **Cosine Similarity (CS, ↑)**

  Why Cosine Similarity? Cosine Similarity evaluates the alignment between high-dimensional feature vectors extracted from synthetic and real datasets.

  Desirable Behavior: Higher Cosine Similarity reflects better feature alignment, demon-

Table 6: Synthetic Medical Data (SMD) Card (Descriptive Section -Part 2)

Synthetic Dataset Training, Validation, and Testing	
Generation Method	Latent Diffusion Model (LDM) with intensity rescaling, augmentation, and noise injection. LDM was trained using CSAW-M dataset.
Validation Process	FID and MS-SSIM metrics validated fidelity and structure similarity.
Reference Dataset (CSAW-M) Information	
Size	Public train: 9,523 images; Public test: 497; Private test: 475.
Source and Origin	Extracted from the CSAW cohort (a collection of millions of mammograms from screening participants aged 40-74 in Stockholm between 2008 and 2015).
Resolution	All images resized to 632×512 and saved in 8-bit PNG format.
Labels and Reference Standard	• **Classes:** Interval cancers, large invasive cancers, composite cancers, and controls. • **Class Distribution:** Interval cancers (35%), Large invasive (25%), Composite (15%), Controls (25%). • **Reference Standard:** Expert annotation process is described in the paper (see citation).
Metadata	Contains clinical endpoints (e.g., cancer type), image acquisition attributes (e.g., laterality, density), and quantitative density measures (e.g., percent density).
Preprocessing and Augmentation	• **Preprocessing:** Horizontal alignment, intensity rescaling, zero-padding, and removal of masking text. • **Augmentation:** Includes rotation, flipping, scaling, brightness adjustment, and noise injection to improve diversity.
Hardware	Hologic devices at the Karolinska breast center.
Known Limitations	Demographic gaps, including underrepresented breast density subgroups. Some metadata fields may be incomplete or ambiguous.
Versioning	Dataset is versioned. Current version: v1.0 (Updated October 2023).
Citation (CSAW-M)	Strand, F., et al. (2021). A population-based mammography dataset for evaluating masking of breast cancer. *Nature Scientific Data*, *8*(1), 140. Available at: https://doi.org/10.1038/s41597-021-00935-2.

strating that the synthetic data captures the same high-level feature patterns as the real dataset.

- **Earth Mover's Distance (EMD, ↓)**
  Why EMD? EMD measures the minimum cost of transforming one probability distribution into another. For this evaluation, it quantifies the distance between feature distributions (VGG16) extracted from synthetic and real datasets.
  Desirable Behavior: Lower EMD values indicate that the synthetic and real feature distributions are closely aligned.

- **Composite Score (CS, ↑)**
  Why Composite Score? The Composite Score aggregates all the normalized metrics into a single value to provide an overall measure of congruence.
  Desirable Behavior: A higher Composite Score indicates overall better alignment across all evaluated metrics.

Normalization for Composite Score Each metric is normalized to ensure comparability:

- **SSIM and Cosine Similarity:** Already in $[0, 1]$.

- **PSNR:** Normalized to $[0, 1]$ using the range $[10, 50]$:

$$PSNR_{Norm} = \frac{PSNR - 10}{50 - 10}$$

- **JSD and EMD:** Inverted to $[0, 1]$ since lower values are desirable:

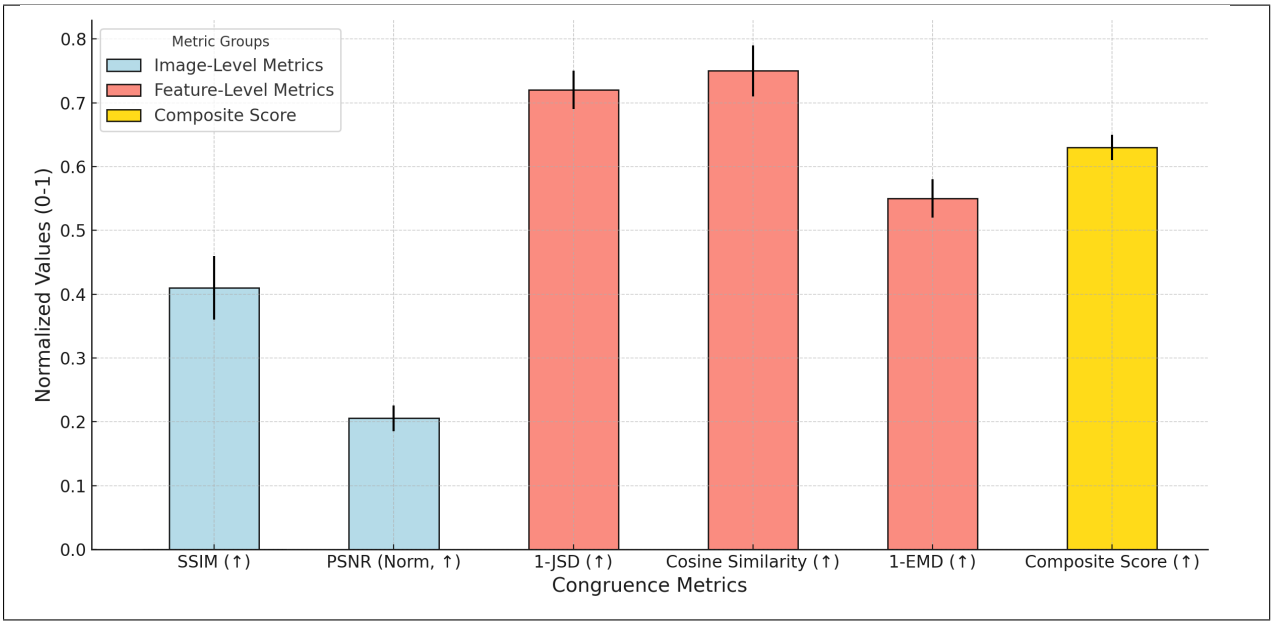
$$JSD_{Norm} = 1 - JSD, \quad EMD_{Norm} = 1 - EMD$$

The Composite Score is computed as:

$$Composite\,Score = \frac{SSIM_{Norm} + PSNR_{Norm} + JSD_{Norm} + CS_{Norm} + EMD_{Norm}}{5}$$

Table 7: Numerical Results of Congruence Metrics for Sinkove. The accompanying figure provides a visual representation of normalized metric values, with error bars indicating the uncertainty (standard deviation). Metrics are grouped into three categories: image-level metrics (light blue), feature-level metrics (salmon), and the composite score (gold).

Metric	RMD (CSAW-M)	SMD (Sinkove)	Interpretation
SSIM (↑)	N/A	0.41	Moderate structural similarity between Sinkove and CSAW-M.
PSNR (↑)	N/A	18.22	Reasonable reconstruction quality.
JSD (↓)	N/A	0.28	Good feature alignment between CSAW-M and Sinkove datasets.
Cosine Similarity (↑)	N/A	0.75	High similarity in feature space.
EMD (↓)	N/A	0.45	Moderate alignment; some differences in distribution tails.
Composite Score (↑)	N/A	0.63	Moderate overall congruence between Sinkove and CSAW-M datasets.



Conclusion: Sinkove exhibits strong feature-level congruence (Cosine Similarity and 1-JSD) with CSAW-M but shows moderate structural similarity (SSIM) and reconstruction quality (PSNR), with an overall composite score of 0.63, indicating room for improvement in pixel-level fidelity.

D.2.2 Coverage

Coverage evaluates the diversity and novelty of the synthetic dataset. This dimension measures whether the synthetic data sufficiently captures the range of variability and patterns present in the real dataset.

Metrics and Justification
We used the following metrics to assess coverage for Sinkove. Each metric captures a specific aspect of diversity and novelty:

- **Variance (↑)**
  Why Variance? Variance measures the spread of the data distribution in feature space.
  Desirable Behavior: Higher variance indicates better representation of diverse patterns.

- **Entropy (↑)**
  Why Entropy? Entropy quantifies the uncertainty or randomness in feature distributions. A higher entropy suggests the dataset captures a richer variety of features.
  Desirable Behavior: Higher entropy values imply better coverage of unique patterns.

- **Convex Hull Volume (↑)**
  Why Convex Hull Volume? Convex Hull Volume measures the feature space occupied by the synthetic dataset. A larger volume indicates broader coverage of the feature space.
  Desirable Behavior: Higher convex hull volume reflects greater diversity.

- **Rarity Score (↑)**
  Why Rarity Score? Rarity Score evaluates whether the synthetic dataset captures rare or unique patterns in the real dataset.
  Desirable Behavior: Higher rarity scores indicate the synthetic dataset represents infrequent patterns effectively.

Normalization for Composite Score Each metric is normalized to ensure comparability:

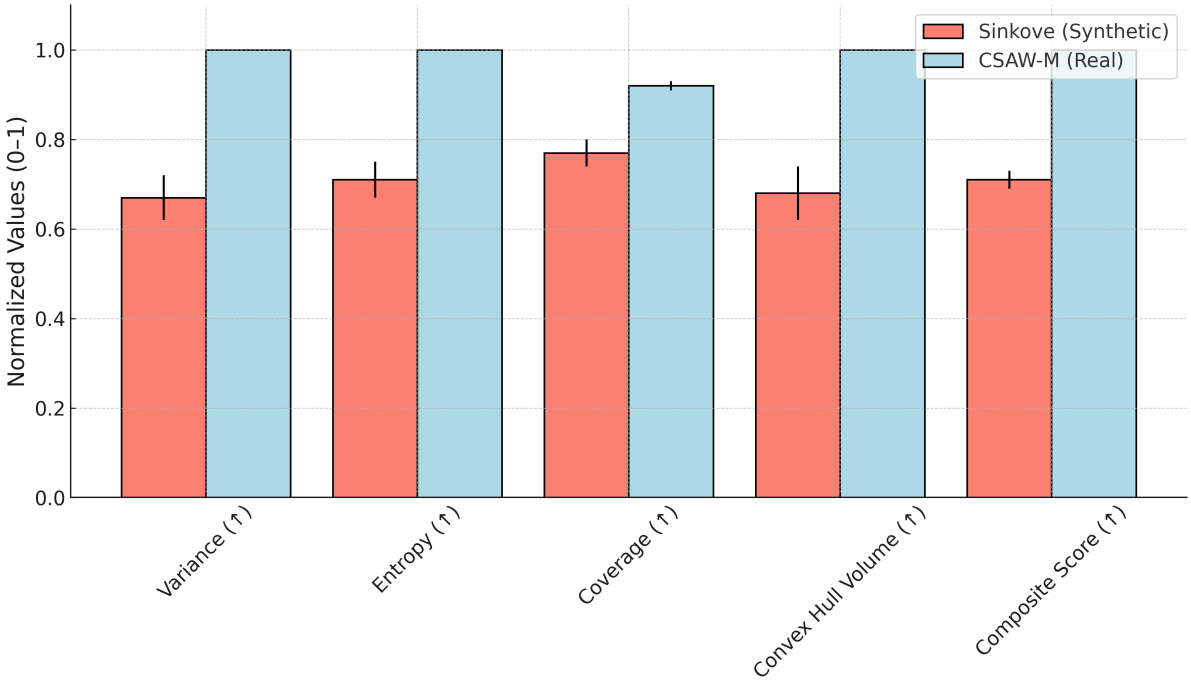
- **Variance, Entropy, and Convex Hull Volume:** Scaled to $[0, 1]$ based on observed ranges.

- **Rarity Score:** Already normalized to $[0, 1]$.

Note on Application These metrics are applied to the feature space extracted from both the synthetic and real datasets using VGG16. Other statistical descriptors or feature extractors can also be used depending on the specific application or data type. Additionally, these metrics can be applied to metadata attributes as features, allowing us to assess whether the synthetic dataset captures the diversity and distribution of metadata present in the real dataset. For example, metadata attributes such as breast density, pathology labels, or imaging device types can be analyzed to ensure the synthetic dataset aligns with the real dataset in terms of subgroup representation and variability.

Table 8: Numerical Results of Coverage Metrics for Sinkove. The accompanying figure provides a visual representation of normalized metric values, with error bars indicating the uncertainty (standard deviation). Metrics are grouped into feature-level metrics (salmon) and the composite score (gold).

Metric	RMD (CSAW-M)	SMD (Sinkove)	Interpretation
Variance (↑)	88.2	59.4	Sinkove shows reduced diversity compared to CSAW-M.
Entropy (↑)	11.2	8.0	Moderate feature diversity in Sinkove but less than CSAW-M.
Convex Hull Volume (↑)	10.2	6.9	Sinkove has a narrower feature space compared to CSAW-M.
Rarity Score (↑)	0.92	0.77	Sinkove captures unique patterns but less effectively than CSAW-M.
Composite Score (↑)	N/A	0.65	Moderate overall coverage of feature space by Sinkove.



Conclusion: Sinkove captures moderate feature space coverage (e.g., variance and entropy) compared to CSAW-M, with a composite score of 0.65. While it represents the major patterns, the reduced convex hull volume and rarity score suggest limited diversity in rare or unique features.

D.2.3 Constraint

The constraint dimension assesses whether the synthetic data complies with predefined rules or constraints essential for maintaining clinical and imaging relevance. For digital mammography, these constraints include matching anatomical noise, quantum noise, and frequency patterns with the reference dataset (CSAW-M). Additionally, clinical relevance is evaluated by ensuring proper breast shapes and the presence of a single nipple per image.

Metrics and Justification

We used the following metrics to assess constraint adherence for Sinkove (synthetic dataset). Each metric captures a specific aspect of constraint satisfaction:

- **Anatomical Noise (β)**
  Why Anatomical Noise? Anatomical noise quantifies the level of anatomical variations

and irregularities in the data.
Desirable Behavior: Similar anatomical noise levels between synthetic and real datasets indicate realistic anatomical structures.

- **Quantum Noise (β)**
  Why Quantum Noise? Quantum noise arises from the stochastic nature of photon interactions during image acquisition.
  Desirable Behavior: Synthetic data should mimic the quantum noise levels in real images to achieve similar imaging characteristics.

- **Low Frequency Energy (↑)**
  Why Low Frequency Energy? Low-frequency energy captures global structural patterns in the image.
  Desirable Behavior: Similar levels of low-frequency energy between synthetic and real datasets ensure comparable global structures.

- **High Frequency Energy (↑)**
  Why High Frequency Energy? High-frequency energy represents fine details and texture in the image.
  Desirable Behavior: Maintaining similar high-frequency energy ensures synthetic data has realistic textures and fine details.

- **Constraint Satisfaction Rate (CSR, ↑)**
  Why CSR? Measures the percentage of synthetic data that meets all predefined constraints (e.g., correct shape or one nipple per image).
  Desirable Behavior: High CSR indicates adherence to constraints.

D.2.4 Completeness

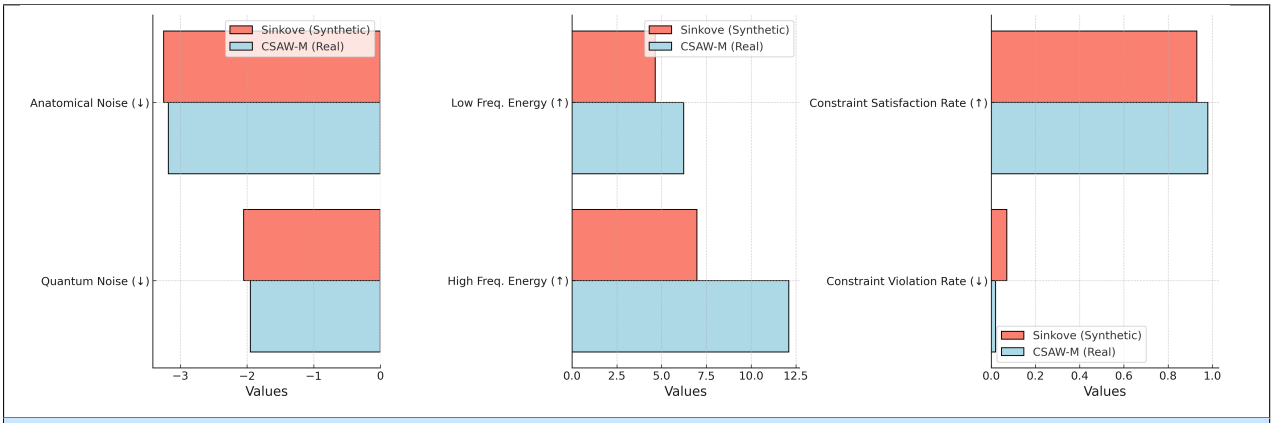
The completeness dimension evaluates whether the synthetic dataset contains all the necessary labels and metadata provided in the reference standard. For digital mammography, the required attributes include both clinical labels and imaging metadata that ensure interpretability and usability.

Required Metadata and Labels The following attributes are required for completeness, based on the reference dataset (CSAW-M):

- **Labels:**
  - **Masking Potential:** Categorized into low, medium, and high masking levels.
  - **Cancer Type:** Includes interval cancers, large invasive cancers, composite cancers, and controls.
  - **Lesion Characteristics:** Such as mass size, shape, and location.
- **Metadata:**
  - **Acquisition Conditions:** Imaging device information, radiation dose, and acquisition time.

Table 9: Numerical Results of Constraint Metrics for Sinkove. The accompanying figure provides a visual representation of metric values.

Metric	RMD (CSAW-M)	SMD (Sinkove)	Interpretation
Anatomical Noise (β)	-3.18	-3.25	Similar noise patterns between Sinkove and CSAW-M.
Quantum Noise (β)	-1.95	-2.05	Slightly higher deviation in Sinkove.
Low Freq. Eng. (↑)	6.24	4.64	Narrower range of low-frequency energy for Sinkove.
High Freq. Eng. (↑)	12.10	6.97	Reduced fine structure details in Sinkove, affecting representation of high-resolution features.
CSR (↑)	0.98	0.93	Good adherence to constraints in Sinkove.
CVR (↓)	0.02	0.07	Slightly higher violations in Sinkove compared to CSAW-M.
Composite Score (↑)	N/A	0.85	Acceptable overall adherence to constraints.



Conclusion: Sinkove demonstrates good adherence to anatomical and imaging constraints overall. However, the frequency range of Sinkove does not fully capture the complete range observed in CSAW-M. Also, some images in the dataset violate predefined constraints, such as incorrect breast shapes and the presence of multiple nipples, as shown in Figure ??.

* **Image Attributes:** Laterality (left or right breast) and breast density categories (e.g., fatty, scattered, hetero, dense).
* **Clinical Endpoints:** Follow-up results, pathological confirmations, and treatment information.

Observed Completeness The Sinkove dataset provides only **masking potential labels**, while other critical labels and metadata are missing. This significantly limits its clinical relevance and interpretability compared to the reference dataset.

Completeness Score The completeness score is computed as follows:

$$\text{Completeness Score} = \frac{\text{Number of Included Attributes}}{\text{Total Required Attributes}}$$

$$\text{Completeness Score} = \frac{1}{10} = 0.1$$

This low score indicates significant gaps in the metadata and labels provided by Sinkove.

Conclusion **Conclusion:** The Sinkove dataset provides masking potential labels but fails to include other essential labels and metadata required for clinical relevance and interpretability. This results in a low completeness score of 0.1, indicating significant room for improvement to match the reference standard (CSAW-M).

D.2.5 Compliance

This section evaluates the synthetic dataset along three dimensions:

- **Completeness:** Whether all required metadata and labels from the reference dataset are provided.
- **Compliance:** Whether the dataset adheres to regulatory and ethical requirements, such as de-identification and standardization.
- **Comprehension:** The clarity and availability of documentation, including dataset description, usage instructions, and metadata definitions.

Completeness Evaluation The completeness score assesses the inclusion of critical labels and metadata. The Sinkove dataset provides masking potential labels but lacks most clinical and acquisition metadata required for interpretability.

Compliance Evaluation Compliance ensures that the dataset aligns with regulatory requirements, including:

- **Standardization:** Data formats (e.g., PNG, DICOM) and labeling conventions.
- **De-identification:** Removal of personal identifiers.
- **Differential Privacy (DP):** Quantified as ε for the Sinkove dataset.

Comprehension Evaluation The comprehension score measures the availability and clarity of the dataset's documentation. Sinkove provides limited documentation, which lacks sufficient clarity on acquisition conditions, metadata, and usage guidelines.

D.2.6 Comprehension

Conclusion **Conclusion:** Sinkove exhibits moderate compliance with regulatory requirements, achieving a compliance score of 0.85. However, the dataset shows significant deficiencies in completeness (0.1) and comprehension (0.5). It lacks most clinical labels, metadata, and sufficient documentation, limiting its usability and interpretability compared to the reference dataset (CSAW-M).

D.2.7 Consistency

Given that the Sinkove dataset has the following intended use: "enhance AI model training for cancer masking in mammography", we need to ensure the quality and uniformity of the data across all masking groups. To achieve this, we conducted a subgroup analysis to evaluate and ensure the consistency of the Sinkove dataset's quality across various groups.

Table 10: Evaluation of Completeness, Compliance, and Comprehension for Sinkove Dataset.

Attribute	RMD (CSAW-M)	SMD (Sinkove)	Interpretation
Completeness			
Masking Potential	✓	✓	Provided in both datasets.
Cancer Type	✓	×	Missing clinical labels.
Lesion Characteristics	✓	×	Mass size, shape, and location not included.
Laterality	✓	×	Left/right breast information missing.
Breast Density	✓	×	Density categories not included.
Follow-up Results	✓	×	No longitudinal clinical endpoints provided.
Pathological Confirmation	✓	×	Pathological data missing.
Treatment Information	✓	×	Relevant clinical treatments not included.
Conclusion: Completeness Score = 0.1 (1 out of 10 attributes provided). Given that only masking labels are included, it might restrict its applicability to a narrow range of tasks and limit its generalizability.			
Compliance			
Standardization	✓	✓	PNG format; adheres to standardization.
De-identification	✓	✓	No personal identifiers found.
Differential Privacy (DP)	✓	$\varepsilon = 0.20$	Provides moderate privacy guarantees.
Conclusion: Compliance Score = 0.85 (3 out of 3 requirements met, $\frac{1+1+0.7}{3}$). The compliance score is calculated based on adherence to three criteria: standardization, de-identification, and differential privacy. The dataset uses the PNG format and meets de-identification requirements in accordance with US HIPAA regulations. Differential privacy is implemented with $\varepsilon = 0.20$, which aligns with the minimum acceptable standards.			
Comprehension			
Usage Guidelines	✓	User Study Score = 3	Users reported moderate difficulty in understanding the usage of the dataset. The score was computed based on a user study where participants rated the clarity of provided usage instructions on a 5-point Likert scale. The study involved 20 participants with backgrounds in medical imaging.
Documentation Clarity	✓	User Study Score = ?	??
Generation Process	✓	User Study Score = 1.5	Users reported difficulty in understanding the generation process. The score was derived from the same user study.
Conclusion: Based on a user study involving 20 participants, Sinkove has a moderate usage score (2.5/5) but ranks low in the generation process (2.0/5). The latter is a common limitation of generative AI models, which often lack interpretability due to their black-box nature.			