APPENDIX 12: Evaluating Fidelity and Utility of Synthetic Health Data

Purpose and scope

This appendix provides guidance on **how to evaluate the fidelity and utility of tabular synthetic health data**, recognising that:

- High statistical resemblance alone does **not** guarantee usefulness for health research or system analysis.
- High utility without appropriate safeguards may indicate elevated **privacy risk**.
- No single metric or score is sufficient; evaluation must be **use-case driven, multi-dimensional, and transparent**.

This appendix applies to **tabular, record-level synthetic health data**, including patient-level, episode-level, or service-level datasets, and is intended to be used alongside:

- **Step 3** (Synthetic data generation and validation)
- **Step 4** (Re-identification risk assessment)
- **Appendix 7** (Privacy evaluation methods)

Core concepts and definitions

- **Fidelity (intrinsic resemblance): Fidelity** refers to the degree to which synthetic data reproduces the **statistical properties and structural relationships** of the source data.

Typical questions addressed by fidelity metrics include:

- Do marginal distributions resemble the source?
- Are dependencies between variables preserved?
- Are conditional relationships (e.g. within subgroups) stable?
- Can synthetic records be statistically distinguished from real ones?

Fidelity metrics are **necessary but not sufficient** for judging suitability of synthetic health data.

- **Utility (extrinsic performance): Utility** refers to whether synthetic data can **support the intended analytical or operational task** with acceptable degradation relative to real data. Utility is **context-dependent** and must be evaluated against a **declared use case**, such as:
- cohort feasibility and planning
- analytic reproducibility
- method or pipeline development
- predictive modelling

Utility evaluation is therefore **task-based**, not purely statistical.

- **Relationship to privacy risk:** In health contexts, fidelity, utility, and privacy are **interdependent**:
  - Very high fidelity may indicate elevated disclosure risk.
  - Privacy-preserving mechanisms (e.g. differential privacy) may reduce utility.
  - Evaluation results must be interpreted **jointly**, not in isolation.

This appendix does not replace privacy assessment (Appendix 7), but **complements it**.

Recommended evaluation pipeline

The Framework recommends a **four-stage evaluation pipeline**, proportionate to the intended use and risk profile.

Stage A: Pre-flight validity and constraint checks

Purpose: ensure basic plausibility before scoring fidelity or utility.

Minimum checks should include:

- value ranges and formats
- impossible or contradictory combinations
- missingness patterns (overall and stratified)
- clinical or coding constraints (where applicable)

Failure at this stage indicates the synthetic data is **not fit for downstream evaluation**.

Stage B: Fidelity evaluation (intrinsic resemblance)

Purpose: assess statistical similarity between real and synthetic data.

At minimum, evaluations should include:

1. univariate distribution similarity
2. multivariate dependency preservation
3. subgroup / conditional fidelity
4. at least one distinguishability test

All fidelity metrics should be reported **overall and stratified** by key clinical or demographic variables.

Stage C: Utility evaluation (use-case driven)

Purpose: assess whether synthetic data supports the **declared purpose**.

Utility evaluation must be anchored to **explicit workloads**, such as:

- predictive modelling (TSTR / TRTS)
- analytic replication (effect estimates, rates, rankings)
- cohort logic and feasibility replication

Synthetic data **must not be described as "high utility" without specifying the task(s) tested**.

Stage D: Decision support and documentation

Purpose: support governance decisions, not to compute a single "quality score".

Outputs should include:

- a multi-criteria scorecard or summary table
- a narrative explanation of strengths, limitations, and caveats
- clear statements of **appropriate and inappropriate uses**

Fidelity metrics: recommended categories

Metric category	What it evaluates	Why it matters in health	Typical tools
Univariate distribution similarity	Marginal distributions per field	Detects obvious distortion, missingness shifts	SDMetrics, SynthCity
Dependency preservation	Correlations / associations	Many clinical inferences rely on relationships	SynthCity, SynthEval
Conditional / subgroup fidelity	Stability within strata (e.g. age, sex, site)	Health utility often fails in rare or minority groups	Custom + SynthCity
Distinguishability	Ability to separate real vs synthetic	Flags over- or under-fitting	SynthCity

Governance note: indistinguishability alone must never be used as evidence of utility or safety.

Utility metrics: recommended categories

Utility category	Evaluation approach	Appropriate when	Notes
Predictive utility	TSTR / TRTS model performance	ML development, benchmarking	Report overall + stratified
Analytic replication	Reproduce target analyses	Research planning, policy	Compare effects, CIs, rankings
Cohort replication	Counts, rates, inclusion logic	Feasibility, service analysis	Highly governance-aligned

Tools and platforms (non-exhaustive)

Open-source evaluation libraries

- **SDMetrics**: model-agnostic fidelity reports
- **SynthCity:** integrated fidelity, utility, detection, privacy benchmarks
- **SynthEval**: structured evaluation framework for tabular data
-

Health-focused benchmarking

- **SynthRO:** prioritised, dashboard-based benchmarking for health datasets
-

Commercial platforms (with evaluation modules)

- YData, MOSTLY AI, Gretel (evaluation outputs must be documented and reproducible)
-

Framework expectation: tool choice must be documented; results must be interpretable without vendor-specific scoring alone.

Persona-aligned minimum evaluation sets

Data Custodian (governance approval):

Minimum required

- Pre-flight validity checks
- Univariate + dependency fidelity (stratified)
- One utility workload aligned to stated purpose
- Explicit linkage to Appendix 7 privacy assessment

Decision question:

Is this synthetic dataset sufficiently faithful and useful *for the stated purpose*, without introducing unacceptable residual risk?

Data Requestor (research / planning user):

Minimum required

- Utility evaluation aligned to their proposed analysis
- Clear limitations on unsupported uses
- Cohort and rate replication where relevant

Decision question:

Can this dataset answer my question reliably, and where could mislead me?

Data Scientist (generator / evaluator):

Minimum required

- Full fidelity suite (including conditional fidelity)
- At least two utility workloads
- Diagnostic plots and failure analysis
- Documentation of trade-offs observed
-

Decision question:

Which generation approach best balances fidelity, utility, and privacy for this context?

Reporting requirements

Every released synthetic dataset should be accompanied by a **Synthetic Data Fidelity & Utility Statement** containing:

1. Declared use case(s)
2. Evaluation methods and tools used
3. Summary of fidelity findings
4. Summary of utility findings
5. Known limitations and exclusions
6. Relationship to privacy assessment (Appendix 7)

This statement is intended to **travel with the dataset** and support responsible reuse.

Key limitations and open issues

- There is **no universal metric set** suitable for all health use cases, however,
- However, it is expected that, as more synthetic data sets are created, some common patterns will emerge.
- Subpopulation fidelity remains the most common failure mode.
- Trade-offs between privacy, fidelity, and utility require explicit judgement, not automation.
- Evaluation practices are evolving; this appendix should be reviewed periodically.

Sample Decision form

Synthetic Health Data — DAC / HREC Decision Form

(Fidelity & Utility Assessment)

Dataset name: _____
Proponent / project: _____
Meeting date: _____

1. Purpose and scope

☐ The intended use of the synthetic dataset is clearly defined
☐ The proposed use aligns with the dataset's stated purpose
☐ Uses that are *not* supported are explicitly documented

Approved use case(s):

2. Fidelity (statistical realism)

☐ Basic validity checks have been performed (no implausible records)
☐ Key patterns and relationships relevant to the use case are preserved
☐ Fidelity has been assessed for important subgroups (e.g. age, sex, ethnicity, site)
☐ Known distortions or limitations are clearly described

Committee note (if any):

3. Utility (fitness for purpose)

☐ Utility has been tested against the *actual intended task*
☐ Results are understandable without technical expertise
☐ Performance is acceptable given the stated purpose and risk profile
☐ Risks of misuse or misinterpretation are documented

Primary utility evidence used:
☐ Analytic replication
☐ Predictive task (TSTR / similar)
☐ Cohort / rate replication
☐ Other: _____

4. Privacy and risk context

☐ Fidelity and utility findings have been reviewed alongside privacy assessment
☐ Any trade-offs between realism and privacy are acknowledged
☐ Residual risks are justified and proportionate to intended use

5. Documentation and transparency

☐ A plain-language Fidelity & Utility Statement is provided
☐ Evaluation methods and assumptions are recorded
☐ Appropriate and inappropriate uses are clearly stated

6. Committee decision

Is the synthetic dataset fit for the stated purpose?

☐ Approved
☐ Approved with conditions
☐ Not approved

Conditions / requirements (if applicable):

Chair / delegate signature: _____ Date: _____

Decision principle (for record)

Approval reflects judgement that the dataset is sufficiently faithful and useful for the stated purpose, with acceptable residual risk, not that it is universally "high quality".