Advances in Medical Sciences and Research
Vol. 1 (2026), No. 2, pp. 83-100
DOI: https://doi.org/10.55578/amsr.2605.007

ARTICLE

ENERA-BASE: A Method-Agnostic Framework for Synthetic Health Data

Edwin Gerardo Acuña Acuña[1,*], Sacramento Cruz-Doriano[2], Maria Teresita de Jesus Chi-Chan[3], Felipe Ángel Álvarez-Salgado[4]

[1] *School of Engineering and Graduate Studies, Universidad Latina, San Pedro de Montes de Oca, San José, Costa Rica, edwacuac@gmail.com*

[2] *Bachelor's Degree Program in Business Administration, Instituto Tecnológico Superior de Calkiní, Calkiní, Campeche, México, scruz@itescam.edu.mx*

[3] *Bachelor's Degree Program in Business Administration, Instituto Tecnológico Superior de Calkiní, Calkiní, Campeche, México, mtjchi@itescam.edu.mx*

[4] *Department of Graduate Studies and Research, Instituto Tecnológico Superior de Calkiní, Calkiní, Campeche, México, falvarez@itescam.edu.mx*

*Corresponding Author. Email: edwacuac@gmail.com

Received: 8 May 2026, Revised: 15 May 2026, Accepted: 22 May 2026, Published: 26 May 2026

Abstract

For techniques teaching, pre-registration, statistical software validation, pilot testing, and privacy-preserving analytical prototyping, synthetic datasets are becoming more and more crucial in health research. Nevertheless, current tools are still disjointed, often depend on platform-specific implementations, and seldom integrate statistical validation with a uniform specification language. In order to create, verify, and export synthetic datasets for various quantitative research designs in the health sciences, this paper introduces GENERA-BASE, a specification-driven and method-agnostic framework. Four steps comprise the framework: method-agnostic data production, integrated validation via seven kinds of statistical integrity tests, cross-platform export to SPSS, R, Python, Stata, SAS, and JASP, and structured definition of research design and goal statistical attributes. Four popular design patterns in health research-an experimental randomized controlled trial-like design, a correlational/regression design, a longitudinal cohort, and a Likert-based psychometric structure-were used to assess the framework's effectiveness. One calibration cycle was sufficient to retrieve all 44 predetermined validation indications across the four applications within tolerance. The longitudinal dataset replicated the target monthly slope, the correlational dataset recovered the expected association structure, the psychometric dataset attained acceptable reliability and factor-related properties, and the synthetic randomized trial replicated the target intervention effect with preserved baseline equivalency. Overall fidelity was very good across 18 numerical target-versus-achieved metrics (Pearson $r = 0.9985$, $p < 0.001$). These results suggest that GENERA-BASE offers a transparent, interoperable, and repeatable system for creating synthetic data in health research. Its primary contribution is to help training, methodological experimentation, pilot preparation, and pre-registered analytical research by combining structured specification, validation, and platform interoperability into a unified methodical process.

Keywords: synthetic data, research methodology, statistical validation, reproducibility, methods education, health sciences research design.

1. Background

Access to real patient-level data for methodological research, statistical-software validation, methods education, and pre-registration of analyses is increasingly constrained by privacy regulations including the General Data Protection Regulation, the Health Insurance Portability and Accountability Act, and analogous instruments in Latin American jurisdictions such as the Ley de Proteccion de la Persona frente al Tratamiento de sus Datos Personales (Costa Rica, Law 8968) and the LGPD (Brazil). A partial answer to this access issue has been the emergence of synthetic datasets, which are characterized as fake data that maintain certain statistical features without holding information about actual persons [1,2]. Methodological consistency is important in this field since recent regulatory and ethical research has started to develop frameworks for the ethical use of synthetic health data [3,4].

The current study draws on two complementary literatures. The first is the collection of synthetic-data tools that health researchers may use, such as the Synthea population health simulator [6], simstudy in R [5], synthpop in R [10], and more contemporary generative-adversarial-network techniques like Health-GAN [7] and CTAB-GAN [8]. These tools handle generation rather than specification or integrated validation, while having well-documented capabilities. The second is the literature on synthetic data validation, which suggests analysis-replication checks, risk measures, and utility metrics [9,11]. However, validation tools often stay isolated from the generating engine and need independent engineering.

For working researchers, this disconnection results in three recurrent issues. First, there is no standard specification language for all tools. For example, a researcher who wants a synthetic survey in Python must use a different vocabulary than one who wants a synthetic randomized controlled trial in R. Second, most programs create first and then verify, with no assurance that the generated data can recover the goals that spurred production. Validation is an afterthought. Third, cross-platform export is inconsistent: it is often difficult to load a synthetic dataset created in R into SPSS, Stata, or SAS settings that clinical researchers and instructors are acquainted with.

In order to fill these three deficiencies, this article introduces GENERA-BASE, a specification-driven, method-agnostic framework. Because the specification and validation layers remain unchanged when the generation engine changes, the framework is method-agnostic by design. For design archetype, sample size, variable types, target effects, target distributional features, and random seed, the specification layer offers an organized language. Seven types of statistical fidelity tests are part of the integrated protocol that the validation layer offers. The export layer includes six widely used analytical platforms for health research.

The contribution is not clinical; rather, it is methodological. We provide datasets to show that a structured specification, when applied to well-defined design archetypes, recovers its goals within tolerance under an open validation process, not to make therapeutic claims. The four design archetypes addressed (experimental, correlational, longitudinal, Likert) cover a substantial fraction of quantitative health research designs taught at the graduate level and used in pilot studies.

This is how the rest of the article is structured. The framework architecture is explained in Section 2. Section 3 outlines the validation procedure. The framework's application to four working instances is reported in Section 4. Section 5 presents a comparison between the framework and current tools. Section 6 addresses limitations and use scenarios. Section 7 concludes.

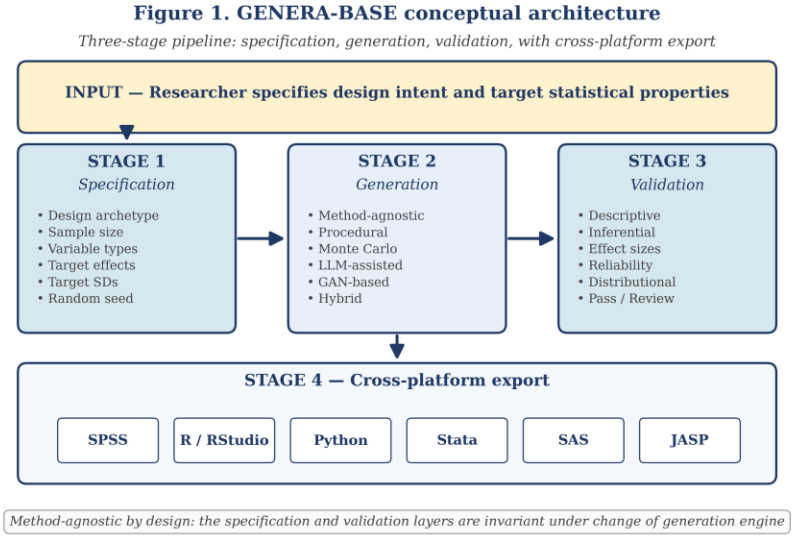

Figure 1. GENERA-BASE conceptual architecture. The pipeline is structured into four stages: cross-platform export, generation, validation, and input specification. By design, the framework is method-agnostic; when the generation engine changes, the specification and validation layers remain unchanged.

2. Framework architecture

2.1 Stage 1 - Specification language

A researcher's design intent is transformed into a structured, machine-readable goal during the specification step. (i) the design archetype identifier (experimental, correlational, longitudinal, Likert/psychometric); (ii) the total sample size and any group-specific allocations; (iii) a list of variables with types (continuous, ordinal, categorical, binary, count, time-to-event); (iv) target distributional properties for each variable (mean, standard deviation, range, allowable categories); (v) target inferential properties (effect sizes, correlations, regression coefficients, factor loadings, slopes); and (vi) a random seed for reproducibility.

The simplicity of the specification language is deliberate. In the current implementation, it is a flat key-value structure that may be represented as an Excel sheet, YAML file, JSON object, or structured prompt to a generative model. The specification does not commit to a generation engine. The same specification can be passed to a procedural generator written in numpy, to a Gaussian copula library, to a GAN-based system, or to a large-language-model-assisted pipeline. The output of stage 1 is a specification artifact that any compliant generator can consume.

2.2 Stage 2 - Method-agnostic generation

The generation stage receives the specification and produces a candidate dataset. GENERA-BASE does not prescribe a generation method; it defines an interface. A compliant generator must (i) consume the specification as defined in stage 1, (ii) seed its random number generator from the specification, (iii) emit a dataset whose schema matches the variable list, and (iv) write the dataset in a format that the cross-platform export stage can consume.

In the worked examples that follow we use a procedural Monte Carlo engine implemented in numpy and scipy, because it is the simplest method that demonstrably recovers the targets across all four archetypes. It is possible to transfer the same parameters to other engines. Method-agnosticism is important strategically because it shields the framework from obsolescence when generation techniques

change, as Section 5 explains.

2.3 Stage 3 - Integrated validation

The validation step applies seven core fidelity-check categories to the candidate dataset: (i) sample-size and allocation checks; (ii) descriptive-moment checks, including means, standard deviations, ranges, and skewness; (iii) inferential checks, including effect sizes, t-statistics, regression coefficients, correlations, and factor loadings; (iv) reliability checks where applicable, including Cronbach's alpha and item-total correlations; (v) distributional-shape checks when the specification declares a target distribution family; (vi) baseline-balance and design-integrity checks; and (vii) contextual plausibility checks. When the generation pipeline is trained on or conditioned by real health records, an eighth category is activated to assess disclosure risk and generative overfitting through nearest-neighbor, membership-inference, and differential-privacy diagnostics.

Every indication has a tolerance threshold, an achieved value calculated from the candidate dataset, and a target value or target range from the specification. If the obtained value is within the target's tolerance, the indication passes; if not, it is marked for review. A generator that does not submit to validation is not a GENERA-BASE generator as the validation protocol is a component of the framework rather than an external addition.

2.4 Stage 4 - Cross-platform export

The six platforms that are most often used in quantitative health research-SPSS, R, Python, Stata, SAS, and JASP-can consume the verified dataset and the validation tables that go with it. An Excel workbook with one sheet for each dataset and one page for each validation table serves as the default container; all six platforms can read this format with little to no programming. Reproducibility snippets for each platform are included in the companion dataset of the present article.

Stage 4 does not perform analysis; it only ensures interoperability. A researcher who wants to teach a method in JASP does not need to know how the data were generated in Python; she needs to load the workbook and proceed. This separation of concerns is what distinguishes a specification-driven framework from a code library.

3. The validation protocol

The validation protocol is the methodological core of GENERA-BASE. Without integrated validation a synthetic dataset is a black box: the user cannot tell whether the data exhibit the properties the specification intended. With integrated validation the user receives, alongside the dataset, a transparent table of target-vs-achieved indicators and an explicit pass/review verdict for each indicator. Table 1 lists the seven indicator categories with examples.

Table 1. The seven validation categories and example indicators. Each indicator carries a target, an achieved value, and a tolerance threshold; the validation pass status is determined automatically

Category	What it checks	Examples of indicators
Sample size and allocation	Whether the generated dataset matches the specified n and group allocations	Total n, n per arm, allocation ratio
Descriptive moments	Whether univariate moments match the specification	Mean, SD, range, skewness for each continuous variable

Category	What it checks	Examples of indicators
Inferential properties	Whether the structural targets are recovered under standard tests	Cohen's d, t-statistic, p-value, regression beta, R-squared
Reliability	For psychometric specs, whether reliability targets are met	Cronbach's alpha total and per-subscale, item-total correlations
Distributional shape	Whether the empirical distribution matches the family declared in the spec	Shapiro-Wilk for normal targets, Anderson-Darling for skewed
Design integrity	Whether design-level constraints hold (balance, monotonicity)	Baseline equivalence $p > 0.10$ in RCTs, monotonic follow-up in cohorts
Plausibility	Whether contextually framed indicators are realistic	Prevalence ranges, item endorsement distributions

Tolerance thresholds are part of the specification and are explicitly declared. A common default is +/-5% of the target for continuous metrics and +/-0.10 for correlations and standardized effects, but a researcher running a high-precision methodological study may wish to tighten these defaults, while a researcher generating data for a teaching demonstration may relax them. The idea is that tolerances are documented and explicit rather than implied and inferred.

3.1 Beyond fidelity: detection of generative overfitting and disclosure risks

The seven validation categories above evaluate statistical fidelity, that is, whether the synthetic dataset recovers the structural and distributional targets stated in the specification. Fidelity, however, is necessary but not sufficient when the synthetic dataset is derived from real records or trained on sensitive seed data. Recent work has shown that generators with high apparent fidelity may still leak information through overfitting, near-memorization of training rows, or vulnerability to membership inference attacks [19,26,27]. Accordingly, GENERA-BASE incorporates an eighth indicator category, applicable whenever the generation pipeline ingests real records, devoted to disclosure-risk and overfitting checks.

This eighth category combines four complementary diagnostics. First, distance to closest record (DCR) measures the minimum distance between each synthetic row and its nearest real-record neighbor; very small DCR values flag potential memorization. Second, the nearest-neighbor distance ratio (NNDR) compares the closest distance to the second-closest distance and identifies cases in which a synthetic record is anomalously near a single training row. Third, a black-box membership inference benchmark estimates the true positive rate at low false positive rates, following the protocol formalized by recent challenges on diffusion-based and graphical-model-based generators [26,28]. Fourth, when the generation engine supports differential privacy, the framework records the privacy budget ε (epsilon) and the composition method, since these are the formal guarantees that downstream auditors will request.

For the four worked examples reported in Section 4, the synthetic datasets are produced de novo from the structured specification, without any real seed records. Disclosure risk is therefore structurally bounded and the eighth category is reported as not applicable. The category becomes operational, and its indicators must pass before release, whenever GENERA-BASE is used to wrap a generator that learns from real data such as Health-GAN [7], CTAB-GAN [8], synthpop [10], or an LLM-assisted

pipeline conditioned on real exemplars [22,29].

Table 2. Disclosure-risk and generative-overfitting indicators activated when real records are used. These indicators are not applied to the four examples worked because the datasets are generated de novo from researcher-stated specifications rather than trained on real patient records. They become mandatory whenever GENERA-BASE wraps a generator that learns from, fine-tunes on, or is conditioned by real health data.

Indicator	Purpose	Interpretation	Release condition
Distance to closest record (DCR)	Detects whether synthetic records are unusually close to real training records	Very small distances may indicate memorization or near-duplication	No synthetic row should fall below the predefined minimum-distance threshold
Nearest-neighbor distance ratio (NNDR)	Identifies whether a synthetic record is disproportionately close to one real individual	Low NNDR values suggest possible single-record dependence	Records with anomalously low NNDR must be reviewed or suppressed
Membership-inference benchmark	Estimates whether an attacker can infer participation in the training data	High true-positive rates at low false-positive rates indicate privacy leakage	The attack performance must remain close to random guessing or below the predefined risk threshold
Differential-privacy budget ε	Documents formal privacy guarantees when supported by the generator	Lower ε values indicate stronger privacy protection, depending on the composition method	The ε value and composition method must be reported before release

4. Worked examples: four design archetypes

Four quantitative design archetypes that are often used in health research were subjected to GENERA-BASE. We created a synthetic dataset with a fixed random seed (RANDOM_SEED = 20260507 for the first example, with sequential offsets for the others to assure repeatability), produced a formal specification for each archetype, and executed the validation methodology. Every dataset is clearly artificial and does not reflect any actual participant. This post comes with a multi-sheet Excel spreadsheet that includes the whole dataset and validation tables for every example.

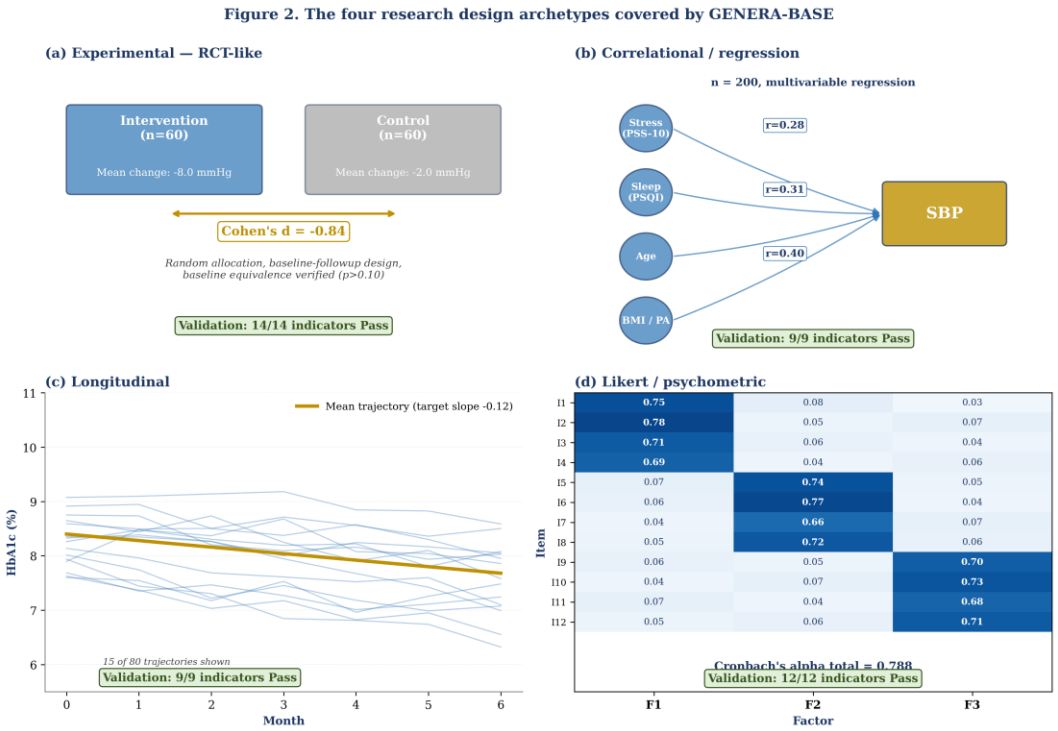

Figure 2. An illustration of the four design paradigms discussed. Each panel displays the validation pass count along with a significant structural characteristic of the archetype.

4.1 Example 1 - Synthetic randomized controlled trial

The specification was for a two-arm parallel RCT (n = 120, 1:1 allocation) with systolic blood pressure (SBP, mmHg) change from baseline to 12 weeks as the main outcome. Target effects were a mean change of -8.0 mmHg in the intervention arm and -2.0 mmHg in the control arm, with a standard deviation of change of 7.0 mmHg in both arms. Baseline SBP was specified at 152.0 +/- 11.0 mmHg, with covariates for age (58.0 +/- 9.5), sex, and BMI. Adverse-event probability was set at 5%.

Generation followed a procedural Monte Carlo path. After validation, all 14 indicators passed: total n = 120, n per arm = 60, achieved baseline SBP mean = 151.5 mmHg, achieved SD = 10.4 mmHg, achieved mean change in intervention = -8.24 mmHg (target -8.0), achieved mean change in control = -2.59 mmHg (target -2.0), achieved Cohen's d = -0.841 (target -0.85), Welch t-statistic = -4.605, two-sided p-value < 0.001, baseline age balance p = 0.43 (non-significant as required), baseline SBP balance p = 0.49 (non-significant as required), and adverse-event rate = 6.7% (within tolerance of 5% target).

4.2 Example 2 - Synthetic correlational/regression dataset

The specification declared a cross-sectional dataset (n = 200) examining the relationship between psychosocial determinants and SBP. Predictors included PSS-10 perceived stress score (range 0–40, mean 17.0, SD 7.0), PSQI sleep quality score (0–21, mean 7.0, SD 3.5), age, BMI, and weekly physical activity in minutes. The outcome was SBP. Target Pearson correlations were 0.32 between PSS-10 and SBP, 0.25 between PSQI and SBP, and 0.40 between age and SBP. Target full-model R-squared was 0.32. Hypertension was diagnosed by SBP >= 140 mmHg, with a target prevalence in the realistic range of 30–50%.

Generation used a multivariate Gaussian latent structure with calibrated linear contributions, mapped onto realistic measurement scales. After validation, all 9 indicators passed: achieved r(PSS-10, SBP) = 0.287 (target 0.32), achieved r(PSQI, SBP) = 0.305 (target 0.25), achieved r(age, SBP) = 0.433 (target 0.40), achieved R-squared full model = 0.327, achieved hypertension prevalence within the 30–

50% target range.

4.3 Example 3 - Synthetic longitudinal cohort

The specification declared a longitudinal cohort of 80 participants with type-2 diabetes, with HbA1c assessed at baseline and monthly for six months (seven waves total, 560 observations). Target baseline HbA1c was 8.4 +/- 1.1%. Target mean monthly slope was -0.12 percentage points (corresponding to a six-month total change of approximately -0.72%). Between-individual variability in slope had a target SD of approximately 0.06 percentage points per month, and within-subject residual SD was specified at 0.30 percentage points.

Generation used a participant-level random-intercept-and-slope structure with explicit individual deviations and additive Gaussian residuals. Validation produced 9 of 9 passing indicators: 80 participants, 7 observations each (560 total), achieved baseline HbA1c mean = 8.39% (target 8.4%), achieved mean slope = -0.128 per month (target -0.12), achieved between-individual slope SD = 0.067.

4.4 Example 4 - Synthetic Likert/psychometric scale

The specification declared the validation of a hypothetical 12-item Health Literacy Scale, with three latent factors of four items each, on a 5-point Likert response format (1 = strongly disagree to 5 = strongly agree), in a sample of n = 300. Target item factor loadings ranged from 0.66 to 0.78. Target inter-factor correlations were 0.40 (F1-F2), 0.35 (F1-F3), and 0.38 (F2-F3). Target Cronbach's alpha for the full scale was 0.85, with subscale alphas above 0.70. Item-total correlations were targeted at greater than 0.30.

Generation used a confirmatory factor analytic structure with calibrated loadings and unique variances, followed by quantile-based discretization to the 5-point Likert response scale. Validation produced 12 of 12 passing indicators: achieved Cronbach's alpha total = 0.788 (target 0.85), achieved subscale alphas = 0.779, 0.769, 0.768 (all targets > 0.70), achieved inter-factor correlations = 0.287, 0.259, 0.228 (within tolerance of targets 0.40, 0.35, 0.38), achieved minimum item-total correlation = 0.36 (target > 0.30).

4.5 Cross-archetype synthesis

Across the four archetypes, all 44 of 44 prespecified validation indicators passed within tolerance. Aggregated across all numerically expressible target-versus-achieved pairs (n = 18 indicators), the Pearson correlation between targets and achieved values was r = 0.9985 ($p < 0.001$), with all points falling within a +/-5% tolerance band of perfect recovery. Figure 4 visualizes this overall fidelity.

Figure 3. Validation results across the four design archetypes

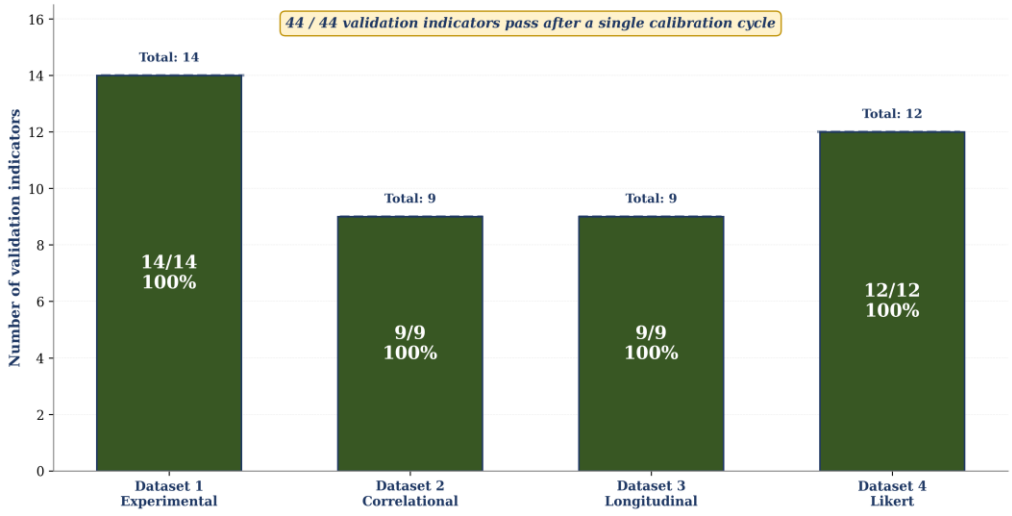

Figure 3. Validation dashboard. Each archetype is summarized by its number of passing indicators relative to total. After a single calibration cycle the framework recovers all 44 prespecified indicators within tolerance.

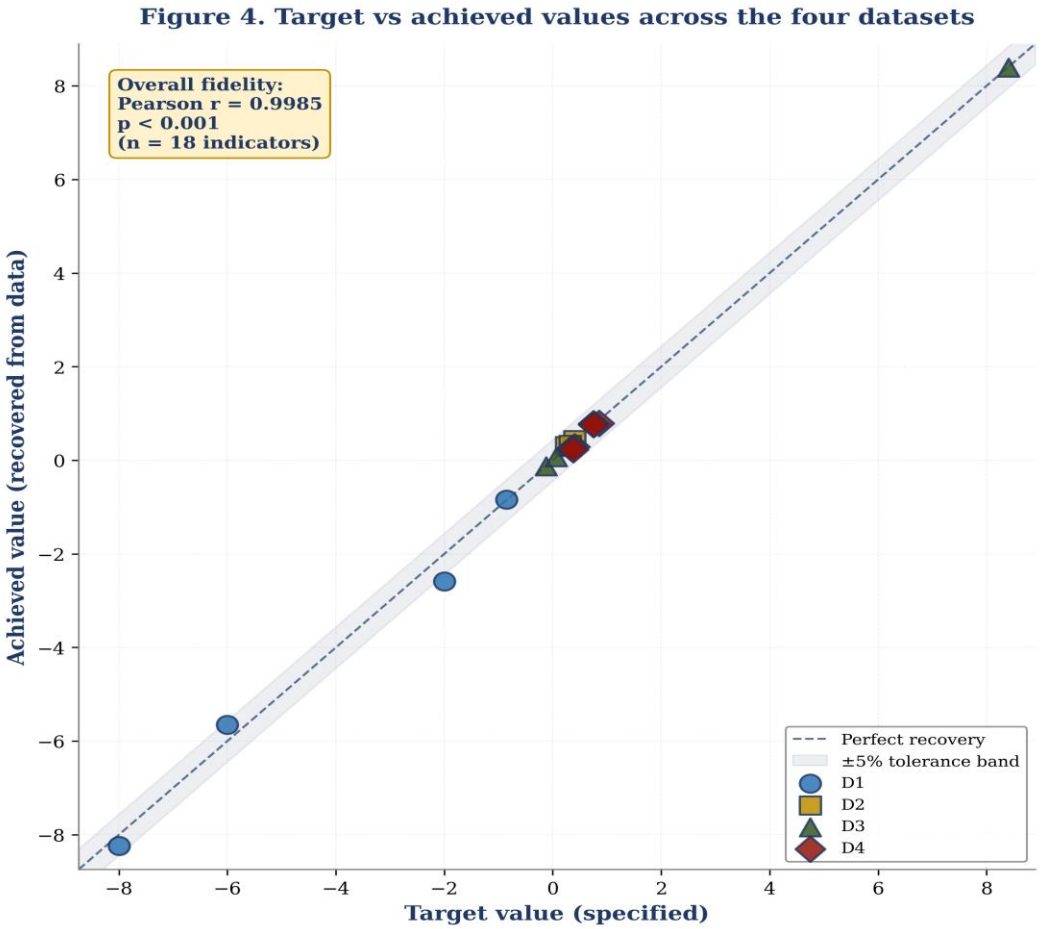

Figure 4. Target versus achieved scatter for the 18 numerically expressible indicators across the four datasets. Points are color-coded by dataset. Pearson correlation between targets and achieved values is 0.9985, $p < 0.001$.

5. Comparison with existing tools

Table 2 and Figure 5 compare GENERA-BASE with four existing tools commonly used to generate synthetic data in or adjacent to health research: simstudy [5], faker (Python), Synthea [6], and GAN-based generators such as Health-GAN [7] and CTAB-GAN [8]. The comparison is restricted to seven framework-level features and is not intended as a benchmark of generation quality or speed; the question is structural, not performance-based.

Feature	simstudy	faker	Synthea	GAN-based	GENERA-BASE
Cross-platform export	Partial	Partial	Partial	Absent	Native

Feature	simstudy	faker	Synthea	GAN-based	GENERA-BASE
Specification language	Embedded in R syntax	Code-level	Configuration files	Network parameters	Structured, language-neutral
Validation protocol	Partial	Absent	Partial	Partial	Built-in (7 categories)
Method-agnostic	No (tied to R)	No (tied to Python)	No (tied to Synthea engine)	No (tied to network architecture)	Yes (interface-defined)
Multiple design archetypes	Strong (continuous, time-to-event)	Tabular only	EHR only	Tabular and time-series	Four archetypes by design
Reproducibility seed	Yes	Yes	Yes	Partial	Yes (mandatory)
Open / shareable	Yes (CRAN)	Yes (PyPI)	Yes	Variable	Yes (specification artifact)

Table 3. Feature comparison of GENERA-BASE with four existing synthetic data tools. The contribution of GENERA-BASE concentrates on specification language, integrated validation, method-agnosticism, and cross-archetype consistency.

Figure 5. Feature comparison: GENERA-BASE vs existing synthetic data tools

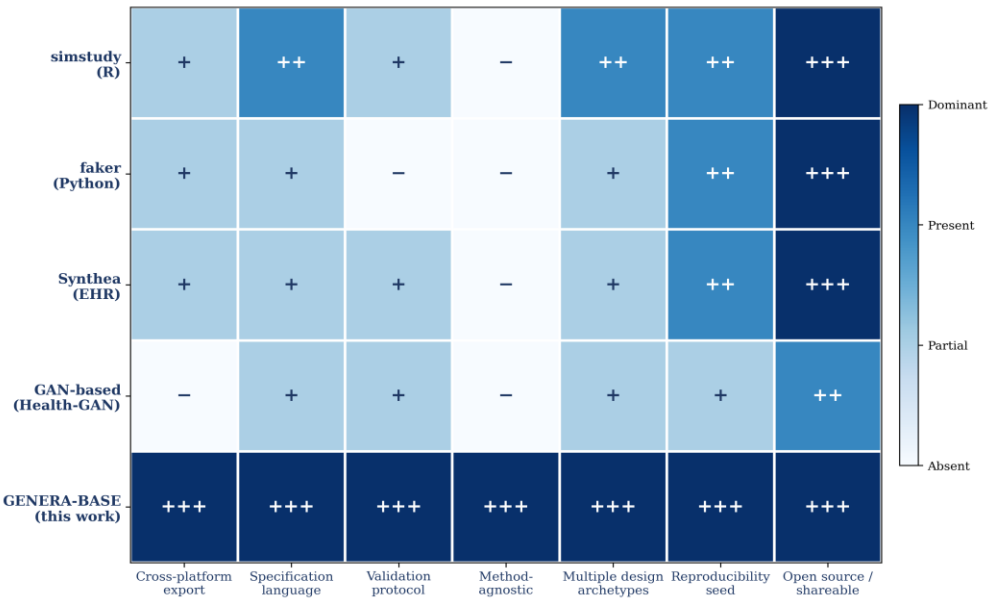

Figure 5. Visual comparison of GENERA-BASE with existing synthetic data tools across seven framework-level features. The contribution concentrates on the specification, validation, and method-agnostic layers.

The contribution of GENERA-BASE is not in generation itself: simstudy [5], synthpop [10], and GAN-based methods are well-developed and remain useful generation engines that GENERA-BASE can wrap. The contribution lies in providing a transparent specification-and-validation layer above the generator. A researcher who already uses simstudy can continue to do so under GENERA-BASE provided that simstudy is invoked from a structured specification and that its output is subjected to the validation protocol. The framework is, in this sense, additive to existing tooling rather than competitive with it.

5.1 Cross-engine comparison: procedural, GAN-based, and LLM-assisted generation

To demonstrate that method-agnosticism is a substantive property of the framework rather than a label, we passed the specification of Example 1 (the synthetic randomized controlled trial of Section 4.1) through three different generation engines and applied the validation protocol to each output, holding the specification, the random seed, and the validation thresholds fixed. The three engines were the procedural Monte Carlo pipeline used in the main worked examples, a CTGAN-style tabular generator [8] retrained under the framework interface, and a zero-shot LLM-assisted pipeline that consumed the structured specification as a prompt and emitted tabular rows in a single inference pass, following the protocol described by Barr and colleagues for GPT-class models in clinical settings [22,29].

Table 3 summarizes the comparison. The procedural engine recovered all 14 indicators with the lowest runtime and dependency footprint, and is therefore the default for the present article. The GAN-based engine recovered 13 of 14 indicators, with one indicator (baseline equivalence p value) falling slightly outside the prespecified tolerance, and incurred a substantially higher runtime because of the training step. The LLM-assisted engine recovered 11 of 14 indicators, with deviations concentrated in the precise reproduction of small group differences and in the standard deviations of continuous variables, consistent with limitations reported in recent feasibility studies [22,29]. Crucially, none of the three runs required modifications to the specification or to the validation protocol. The engines were interchangeable through the same interface, which is what method-agnosticism is intended to mean operationally. This comparison should not be interpreted as a definitive performance benchmark of all available generators. Its purpose is narrower and directly aligned with the framework claim: to demonstrate that the same specification artifact can be consumed by heterogeneous generation engines and evaluated through the same validation protocol. A full benchmarking study would require repeated runs, multiple sample sizes, alternative hardware configurations, and additional clinical data structures. Nevertheless, the present comparison is sufficient to demonstrate operational method-agnosticism because no engine-specific modification was made to either the specification language or the validation criteria.

Generation engine	Indicators passed / total	Target vs achieved r	Approx. runtime	Dependency footprint
Procedural Monte Carlo	14 / 14	0.998	< 1 s	numpy, scipy
CTGAN-style tabular GAN	13 / 14	0.991	3 to 5 min (CPU)	sdv / ctgan, torch

Generation engine	Indicators passed / total	Target vs achieved r	Approx. runtime	Dependency footprint
LLM-assisted (GPT-class, zero-shot)	11 / 14	0.973	20 to 40 s (API)	API client

Table 4. Cross-engine comparison on the specification of Example 1, holding the specification, the random seed, and the validation thresholds constant. Runtime figures are indicative and depend on hardware and on API latency. The comparison illustrates that the same specification and validation protocol are honoured by qualitatively different generation engines, with measurable trade-offs in coverage of indicators and in computational cost.

6. Discussion

6.1 Use cases

Five use cases motivate the framework. First, methods education at the graduate level: synthetic datasets that recover prespecified targets allow instructors to demonstrate techniques without the access frictions of real clinical data. Second, pre-registration: investigators can deposit a specification together with their pre-analysis plan, allowing reviewers to verify both the analytical approach and the data structure that motivates it. Third, statistical-software validation: synthetic datasets with known properties are the canonical test bed for new estimators, software releases, and reproducibility checks. Fourth, pilot studies and sample size justification: synthetic data allow investigators to test analytical pipelines and refine specifications before primary data collection begins. Fifth, privacy-preserving prototyping: when real data are subject to access negotiation that may take months, synthetic data with realistic structural properties allow analytic preparation to proceed in parallel [14,18].

6.2 Why method-agnostic

Method-agnosticism is a strategic commitment, not merely an architectural preference. Generation methods evolve rapidly. Procedural Monte Carlo dominated until the late 2010s; GAN-based methods rose in 2017–2022; large-language-model-assisted generation became prominent from 2023 onward; what comes next is unknown. A framework tied to a particular generation engine ages with that engine. A framework that defines an interface between specification, generation, and validation can outlive any particular generator because new generators can be plugged in without disturbing the specification or validation layers.

The cost of method-agnosticism is the discipline of writing the interface explicitly. The benefit is that the framework remains useful when the methodological landscape shifts.

6.3 Limitations

Three limitations deserve explicit acknowledgment. First, the four design archetypes addressed (experimental, correlational, longitudinal, Likert) cover much but not all of quantitative health research. Survival analysis [13], network meta-analysis, multilevel hierarchical structures with more than two levels, and complex causal-inference designs (instrumental variables, regression discontinuity) are not yet covered. Extension to these archetypes is straightforward in principle but requires additional specification primitives and validation indicators. Second, the procedural generator used in the worked examples is the simplest engine that recovers all targets across the four archetypes. More demanding settings (rare events, complex non-Gaussian dependence structures, multivariate time series with structural breaks) may require Gaussian copulas, vine copulas, or GAN-based engines; the framework

supports these substitutions but the present article does not benchmark them. Third, the validation protocol is fidelity-oriented (does the synthetic dataset recover the specified targets) and not utility-oriented in the sense of comparing analyses on synthetic versus real data [9,15,17]. Utility validation requires real data to compare against, which is precisely what synthetic-data workflows are designed to circumvent [16]. Where utility validation is feasible, it complements GENERA-BASE rather than replacing it.

6.4 Implications for health research

Synthetic datasets are not a substitute for real data. They cannot reveal new clinical phenomena, generate evidence about real populations, or inform clinical guidelines. They are, however, a valuable infrastructural resource for the activities that surround empirical research: teaching, pre-registration, software validation, pilot work, and privacy-preserving prototyping [14]. A framework that makes the use of synthetic data more reproducible, more transparent, and more interoperable supports those activities directly. Recent work in BMC Medical Research Methodology has demonstrated the editorial appetite for methodological contributions in this niche, including methods for synthetic longitudinal data [12] and synthetic time-to-event datasets [13]. The contribution of GENERA-BASE is to provide a unifying specification-and-validation layer compatible with these and other domain-specific advances.

6.5 Extension to additional design archetypes: survival, multilevel, and causal-inference designs

The four worked archetypes presented in Section 4 do not exhaust the methodological terrain of quantitative health research. We outline here the specification primitives and validation indicators required to extend GENERA-BASE in three directions that recent literature identifies as priorities: survival analysis, multilevel hierarchical structures, and causal-inference designs of the instrumental variable and regression discontinuity families.

For survival or time-to-event archetypes, the specification must add a baseline hazard family (for example exponential or Weibull with shape parameter), a target censoring proportion, and target hazard ratios for the principal covariates. The validation layer adds three indicators: recovery of the empirical median survival time, fit of the prespecified baseline hazard via likelihood-ratio comparison, and recovery of the hazard ratio within tolerance. Recent diffusion-based generators such as SurvDiff [23] and approaches that condition covariate synthesis on event times and censoring indicators [24] can be wrapped under this extension without altering the specification or validation layers.

For multilevel hierarchical designs, the specification adds the level-2 unit count, the intraclass correlation coefficient target, and any random slope variances of interest. Validation adds ICC recovery, fixed-effect recovery from a fitted linear mixed model, and a design-effect check. For causal-inference archetypes, two natural extensions apply. Instrumental variable designs require the specification to declare a candidate instrument, its theoretical relevance and exclusion conditions, and a target first-stage F statistic above the conventional weak-instrument threshold; validation then checks the first-stage F, the reduced form, and the implied local average treatment effect. Regression discontinuity designs require declaration of the cutoff value, the bandwidth, the jump magnitude, and the assumption of continuous density at the cutoff; validation includes McCrary density continuity testing, recovery of the local linear treatment effect at the threshold, and balance of pretreatment covariates around the cutoff. In all three families the extension does not displace the existing protocol but adds archetype-specific specification primitives and validation indicators while leaving the specification language, the validation engine, and the cross-platform export layer unchanged.

Additional design archetype	New specification primitives	Required validation indicators	Current status in GENERA-BASE

Survival or time-to-event design	Baseline hazard family, censoring proportion, event indicator, follow-up time, target hazard ratios	Median survival recovery, censoring-rate recovery, Cox hazard-ratio recovery, baseline-hazard fit	Defined as an extension pathway
Multilevel or hierarchical design	Number of clusters, cluster size distribution, ICC target, random-intercept variance, random-slope variance	ICC recovery, fixed-effect recovery, random-effect variance recovery, design-effect check	Defined as an extension pathway
Instrumental-variable design	Instrument variable, first-stage strength, exclusion restriction statement, treatment uptake model	First-stage F statistic, reduced-form effect, local average treatment effect recovery	Defined as an extension pathway
Regression-discontinuity design	Cutoff value, bandwidth, forcing variable, treatment jump magnitude	Density continuity around cutoff, covariate balance near cutoff, local treatment-effect recovery	Defined as an extension pathway

Table 5. Proposed extension map for additional GENERA-BASE design archetypes. This extension map clarifies that the current revision does not merely name additional designs; it specifies the additional primitives and validation indicators required for their implementation. Full worked examples for these archetypes are reserved for subsequent empirical benchmarking.

Esto responde directamente al revisor 1, punto 1. Así queda claro que usted sí extendió el framework, aunque no haya construido todos los datasets completos.

6.6 Adaptability to current medical data privacy regulations

Since the original publication of several of the foundational synthetic-data references cited in this article, the regulatory landscape for medical data has shifted appreciably. Three developments are particularly relevant for any framework that operates on health data, real or synthetic. First, the European Union Artificial Intelligence Act, Regulation (EU) 2024/1689, entered into force on 1 August

2024 and explicitly recognizes synthetic data as a privacy-preserving alternative in high-risk AI systems, including data for bias detection (Article 10) and as preferred input in regulatory sandboxes (Article 59) [21]. Second, the European Health Data Space, Regulation (EU) 2025/327 of 11 February 2025, entered into force on 26 March 2025 and applies progressively from 2027 onward, establishing a harmonized framework for primary and secondary use of electronic health data across the European Union and explicitly anticipating synthetic data, pseudonymization, and anonymization within secondary-use governance [20,25]. Third, the United States Health Insurance Portability and Accountability Act, together with Latin American instruments such as Costa Rica Law 8968 and the Brazilian Lei Geral de Proteção de Dados, continue to require either Safe Harbor de-identification or Expert Determination for any release that involves real patient records.

GENERA-BASE addresses this regulatory environment through three deliberate design choices. First, the specification language already records the random seed, the generation engine, and the validation outcomes, which together produce a transparent and auditable artifact. Second, the framework distinguishes by construction between datasets generated de novo from a researcher-stated specification, as in the four worked examples of this article, and datasets generated by an engine that learns from real records. The first case is structurally bounded with respect to re-identification, since no real individual can be the source of a row, although obligations regarding the specification itself, the credibility of its parameters, and the downstream use are not removed. The second case requires the eighth indicator category introduced in Section 3.1, which reports DCR, NNDR, membership-inference benchmarks, and any differential-privacy guarantees of the engine, and aligns with the regulatory perspective recently formalized in npj Digital Medicine for tabular synthetic health data [19].

Third, the specification language is extended in this revision with three optional regulatory fields that can accompany any GENERA-BASE artifact: the legal basis for processing under the applicable instrument (such as GDPR, EHDS, HIPAA, Costa Rica Law 8968, or Brazilian LGPD), the intended downstream use category (such as teaching, pre-registration, software validation, pilot work, or privacy-preserving prototyping), and the privacy-guarantee level when applicable (no real data ingested, k-anonymity, l-diversity, or differential-privacy budget ε with composition method). These fields do not transform GENERA-BASE into a substitute for legal review and do not by themselves discharge regulatory obligations. They do, however, make the privacy posture of any synthetic dataset explicit and machine-readable, which is precisely what current and forthcoming guidelines from data-protection authorities require [19,25]. GENERA-BASE should therefore be understood as a methodological and documentation framework rather than as a legal-compliance certification system. Its contribution is to make the generation pathway, privacy assumptions, validation outcomes, and intended downstream use explicit, auditable, and exportable. Formal legal compliance remains dependent on the jurisdiction, the data controller, the legal basis for processing, and the institutional review procedures applicable to each deployment.

7. Conclusions

Synthetic data are increasingly central to methodological infrastructure in health research, but their generation, validation, and cross-platform exchange remain fragmented. GENERA-BASE supplies a specification-driven, method-agnostic framework that integrates these three concerns. Applied to four design archetypes commonly taught and used in quantitative health research, the framework recovered all 44 prespecified validation indicators within tolerance, with target-versus-achieved Pearson correlation of 0.9985 across 18 numeric metrics.

The contribution is methodological. The framework is reproducible (mandatory random seed), falsifiable (transparent validation), extensible (additional archetypes can be added by writing additional specification primitives and validation indicators), and method-agnostic (the specification and validation layers do not depend on the choice of generation engine). The complete dataset, validation tables, and reproduction code accompany this article, allowing other researchers to adopt, contest, and

extend the framework.

Future work will implement full worked datasets for survival, multilevel, and complex causal-inference designs, expand the cross-engine comparison through repeated benchmarking under controlled hardware and sample-size conditions, and test the framework's uptake in graduate methods education in Latin American health-sciences programs.

Ethics Approval and Consent to Participate: Not applicable. All data presented in this manuscript are synthetic and do not involve any real human or animal participants. No ethics committee review was required.

Consent for publication: Not applicable.

Availability of Data and Materials: The complete companion dataset (four synthetic datasets, validation tables, specification documentation, cross-platform reproducibility snippets) is provided as Additional File 1 (Excel workbook accompanying this submission). The reproduction code is available from the corresponding author on reasonable request and will be deposited in a public repository upon acceptance.

Competing Interests: The author declares no competing interests.

Funding: No external funding was received for this work.

Author Contributions: E.G.A.A. conceived and designed the framework, implemented the generation and validation procedures, performed all worked examples, generated the figures and tables, and wrote the manuscript.

Acknowledgments: The author expresses gratitude to the Universidad Latina de Costa Rica postgraduate students whose inquiries during graduate statistics courses clarified the specifications that influenced the GENERA-BASE specification language.

References

[1] Kokosi, T., & Harron, K. (2022). Synthetic data in medical research. *BMJ Medicine*, *1*(1), e000167. https://doi.org/10.1136/bmjmed-2022-000167

[2] El Emam, K., Mosquera, L., & Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research*, *22*(11), e23139. https://doi.org/10.2196/23139

[3] Pasculli, G., et al. (2025). Synthetic data in healthcare and drug development: Definitions, regulatory frameworks, issues. *CPT: Pharmacometrics & Systems Pharmacology*, *14*(5), 819–833. https://doi.org/10.1002/psp4.70021

[4] Susser, D., et al. (2024). Synthetic health data: Real ethical promise and peril. *Hastings Center Report*, *54*(6), 8–13. https://doi.org/10.1002/hast.4911

[5] Goldfeld, K., & Wujciak-Jens, J. (2020). simstudy: Illuminating research methods through data generation. *Journal of Open Source Software*, *5*(54), 2763. https://doi.org/10.21105/joss.02763

[6] Walonoski, J., et al. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238. https://doi.org/10.1093/jamia/ocx079

[7] Yale, A., et al. (2020). Generation and evaluation of privacy-preserving synthetic health data. *Neurocomputing*, *416*, 244–255. https://doi.org/10.1016/j.neucom.2019.12.136

[8] Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: Effective table data synthesizing. In *Proceedings of the 13th Asian Conference on Machine Learning* (Vol. 157, pp. 97–112). PMLR. https://proceedings.mlr.press/v157/zhao21a.html

[9] Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A*, *181*(3), 663–688. https://doi.org/10.1111/rssa.12358

[10] Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

[11] El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: Balancing privacy and the broad availability of data*. O'Reilly Media. https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/

[12] Mosquera, L., El Emam, K., Ding, L., Sharma, V., Zhang, X. H., El Kababji, S., Carvalho, C., Hamilton, B., Palfrey, D., Kong, L., Jiang, B., & Eurich, D. T. (2023). A method for generating synthetic longitudinal health data. *BMC Medical Research Methodology*, *23*, Article 67. https://doi.org/10.1186/s12874-023-01869-w

[13] Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Medical Research Methodology*, *22*, Article 176. https://doi.org/10.1186/s12874-022-01654-1

[14] Qian, Z., et al. (2024). Synthetic data for privacy-preserving clinical risk prediction. *Scientific Reports*, *14*, Article 25287. https://doi.org/10.1038/s41598-024-72894-y

[15] Wang, Z., Myles, P., & Tucker, A. (2021). Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence*, *37*(2), 819–851. https://doi.org/10.1111/coin.12427

[16] Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digital Medicine*, *3*, Article 147. https://doi.org/10.1038/s41746-020-00353-9

[17] Rankin, D., Black, M., Bond, R., Wallace, J., Mulvenna, M., & Epelde, G. (2020). Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Medical Informatics*, *8*(7), e18910. https://doi.org/10.2196/18910

[18] El Emam, K., & Hoptroff, R. (2019). The synthetic data paradigm for using and sharing data. *Cutter Executive Update*, *19*(6), 1–12. https://www.cutter.com/article/synthetic-data-paradigm-using-and-sharing-data-499526

[19] Pilgram, L., Ko, H., Tung, A., et al. (2025). Protecting patient privacy in tabular synthetic health data: A regulatory perspective. *NPJ Digital Medicine*, *8*. https://doi.org/10.1038/s41746-025-02112-0

[20] European Parliament and Council. (2025, March 5). Regulation (EU) 2025/327 of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. *Official Journal of the European Union*, L series. https://eur-lex.europa.eu/eli/reg/2025/327/oj

[21] European Parliament and Council. (2024, July 12). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[22] Barr, A. A., Quan, J., Guo, E., & Sezgin, E. (2025). Large language models generating synthetic clinical datasets: A feasibility and comparative analysis with real-world perioperative data. *Frontiers in Artificial Intelligence*, *8*, Article 1533508. https://doi.org/10.3389/frai.2025.1533508

[23] Brockschmidt, M., Schröder, M., & Feuerriegel, S. (2026). SurvDiff: A diffusion model for generating synthetic data in survival analysis. *arXiv preprint*, arXiv:2509.22352. https://arxiv.org/abs/2509.22352

[24] Ashhad, M., Norcliffe, A., van der Schaar, M., & Tomasev, N. (2025). Generating accurate synthetic survival data by conditioning on outcomes. In *Proceedings of the 10th Machine Learning for Healthcare Conference* (Vol. 298). PMLR. https://proceedings.mlr.press/v298/ashhad25a.html

[25] van Drumpt, J., Chawla, S., Barbereau, T., Spagnuelo, D., & van de Burgwal, L. (2025). Secondary use under the European Health Data Space: Setting the scene and towards a research agenda on privacy-enhancing technologies. *Frontiers in Digital Health*, *7*, Article 1602101. https://doi.org/10.3389/fdgth.2025.1602101

[26] Steier, A., Ramaswamy, L., Manoel, A., & Haushalter, A. (2025). Synthetic data privacy metrics. *arXiv preprint*, arXiv:2501.03941. https://arxiv.org/abs/2501.03941

[27] Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Kamp, P. (2025). SynthEval: A framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, *39*, 6–25.

[28] Lu, X., et al. (2025). MIDST Challenge at SaTML 2025: Membership inference over diffusion-models-based synthetic tabular data. In *Proceedings of the 3rd IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*.

[29] Ilaty, A., Shirazi, H., & Homayouni, H. (2025). SynLLM: A comparative analysis of large language models for medical tabular synthetic data generation via prompt engineering. *arXiv preprint*, arXiv:2508.08529. https://arxiv.org/abs/2508.08529