

Original Research

Evaluation of temporal preservation in synthetic longitudinal patient data

Katariina Perkonoja^{a,b}, Parisa Movahedi^b, Antti Airola^b, Kari Auranen^{a,c}, Joni Virta^a^a Department of Mathematics and Statistics, University of Turku, Turku, Finland^b Department of Computing, University of Turku, Turku, Finland^c Department of Clinical Medicine, University of Turku, Turku, Finland

ARTICLE INFO

Dataset link: https://github.com/Nic5472K/ScientificData2021_HealthGym, https://github.com/btheodorou99/HALO_Inpatient, <https://github.com/Katskl/Temporal-preservation-metrics>

Keywords:

Synthetic data generation
Longitudinal patient data
Temporal preservation
Resemblance evaluation
Fidelity metrics

ABSTRACT

Objective: This study introduces a set of metrics for evaluating temporal preservation in synthetic longitudinal patient data, defined as artificially generated data that mimic real patients' repeated measurements over time. **Methods:** The proposed metrics assess how synthetic data reproduce key temporal characteristics, categorized into marginal, covariance, individual-level and measurement structures.

Results: Strong marginal-level resemblance may be observed even when the covariance structures and individual trajectories are substantially different. Temporal preservation is influenced by factors such as original data quality, measurement frequency, and preprocessing strategies, including binning, variable encoding and precision. Variables with sparse or highly irregular measurement times provide limited information for learning temporal dependencies, yielding reduced resemblance between the synthetic and original data.

Conclusion: No single metric adequately captures temporal preservation; instead, a multidimensional evaluation across all characteristics provides a more comprehensive assessment of synthetic data quality. Overall, the proposed metrics elucidate how and why temporal structures are preserved or degraded, enabling more reliable evaluation and improvement of generative models and supporting the creation of temporally realistic synthetic longitudinal patient data.

1. Introduction

Electronic health records (EHRs) are the primary repository of patient data, integrating treatment-related information across health information systems. Although primarily collected for clinical care, EHR data are increasingly used for *secondary purposes*, such as research, development, and education. This expanded use raises privacy concerns and has led to stricter regulation, motivating synthetic data generation (SDG) as a promising solution for improving accessibility while reducing privacy risks. SDG aims to generate data that resemble the original data closely enough to support similar analyses without revealing identifiable information. Interest in SDG has grown rapidly [1–6], and its importance is increasingly recognized at the policy level. For example, synthetic data are envisioned as a key enabler of secure secondary use within the European Health Data Space [7].

Despite this interest, evaluating SDG methods in terms of both data quality and privacy protection remains a nontrivial challenge. Evaluation is commonly organized into up to four domains: *resemblance* (or *fidelity*), *utility*, *privacy* [1–3,5], and *fairness* [8–10]. Within resemblance evaluation, longitudinal patient data (LPD) present unique challenges. LPD, structured tabular data combining static variables

(e.g., demographics) and time-varying measurements, provide a time-sequenced view of patient health. The value of LPD is fundamentally rooted in this temporal dimension, as it enables the separation of between-individual and within-individual variability, a distinction that cross-sectional data cannot achieve [11]. Maintaining this structure is critical; if these dependence structures are not preserved during the SDG process, the resemblance of the data is compromised, limiting the inferences that the longitudinal design can support.

Despite its importance for analytical validity, resemblance of the temporal dimension is frequently insufficiently addressed. A recent systematic review identified that 30% of studies focusing on synthetic longitudinal data generation did not inspect whether temporality was preserved [1]. Furthermore, many studies rely on aggregating metrics over time into a single summary statistic [1], which may obfuscate the underlying distributional structure. Existing evaluation frameworks [12–16] also rarely include metrics that assess temporal preservation. This gap is particularly problematic for *unbalanced* LPD, where the number and timing of measurements vary across patients [17]. Common in secondary-use settings, this unbalanced structure prevents direct comparison of temporal patterns and measurement distributions, further complicating resemblance evaluation.

* Corresponding author at: Department of Mathematics and Statistics, University of Turku, Turku, Finland.

E-mail address: kakype@utu.fi (K. Perkonoja).

<https://doi.org/10.1016/j.jbi.2026.105074>

Received 7 March 2026; Received in revised form 8 June 2026; Accepted 23 June 2026

Available online 29 June 2026

1532-0464/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

To address these gaps, we propose a formalized methodological framework for assessing *temporal preservation*, that is, how well synthetic data reproduce the evolution of variables over time and the distribution of measurements in both timing and frequency. Our framework uses kernel smoothing to handle the unbalanced structure of LPD by aligning comparisons at a coherent set of time points through interpolation. Grounded in explanatory data analysis, it adapts existing metrics for longitudinal evaluation and proposes novel metrics to assess temporal preservation across four key characteristics of longitudinal data: marginal, covariance, individual, and measurement structure, yielding a more complete representation of temporal fidelity. The remainder of this paper is organized as follows: Section 2 reviews existing approaches and their limitations; Section 3.1 introduces the mathematical notation and formalizes univariate temporal preservation; Section 3.2 presents the proposed metrics; and Sections 4 and 5 detail their application and practical implications, while Section 6 concludes this work.

Statement of significance

Problem or issue: Robust evaluation of temporal preservation in synthetic longitudinal patient data, particularly unbalanced data with irregular measurement timing, is lacking.

What is already known: Synthetic data evaluation typically focuses on marginal resemblance, utility, and privacy. Existing temporal metrics often assume balanced data, rely on strong modeling assumptions, aggregate results into opaque summary scores, or ignore measurement timing and subject-level trajectories.

What this paper adds: We introduce a multidimensional set of interpretable metrics assessing marginal, covariance, individual-level, and measurement structures in unbalanced synthetic LPD. Using flexible, non-parametric estimators, the framework clarifies how temporal dependencies are preserved or degraded. An open-source R implementation and empirical illustration on real-world data are provided.

Who would benefit from the new knowledge in this paper: Researchers developing or evaluating synthetic data generators, health data scientists using synthetic LPD, methodologists, and policymakers concerned with secure secondary use of electronic health records.

2. Related work

In an effort to standardize assessment, several comprehensive evaluation frameworks have been recently developed to assess synthetic health data across fidelity, utility, and privacy [12–16]. While these frameworks provide multi-dimensional evaluations, they are primarily designed for cross-sectional datasets. When applied to longitudinal data, these frameworks typically treat time as just another static feature or require researchers to “flatten” temporal sequences into fixed-width prior to evaluation. This workaround fundamentally alters the data structure. Flattening assumes aligned, evenly spaced observations, which artificially removes the unbalanced structure and sparsity inherent to real-world LPD. To avoid the structural distortions caused by flattening, and because these frameworks lack dedicated temporal metrics, authors typically construct ad-hoc metrics specific to their individual studies. Drawing from a recent systematic review [1], these specific metrics generally fall into three methodological paradigms, each of which still presents critical limitations for unbalanced LPD. We detail these paradigms next.

Descriptive and dependency statistics. The most common approach relies on marginal summaries, autocorrelations, Pearson correlations, and transition probabilities [18–29]. While computationally straightforward, these metrics fundamentally struggle with unbalanced LPD. They predominantly assume evenly spaced observations, necessitating missing value imputation that introduces artificial biases into the evaluation. Furthermore, these methods typically aggregate temporal

dynamics into a single overarching score. While single scores simplify method comparison, they critically obfuscate underlying distributional structures and non-linear dependencies. To overcome this, we adapt these foundational metrics using kernel smoothing and the *variogram* [30]. This enables assumption-free assessment of unbalanced data while preserving visual distributional nuance rather than reducing fidelity to a single summary score (Sections 3.2.1, 3.2.2, 3.2.4).

Parametric and model-based approaches. A second paradigm employs model-based evaluations, such as conditional dependency metrics [31, 32] and trend-residual decomposition [33,34]. These approaches capture sequential dependence by fitting parametric models to patient trajectories. However, their critical flaw in secondary-use settings is the reliance on rigid model assumptions. If the chosen evaluation model misspecifies the true underlying data generation process, a frequent issue with highly irregular EHR data, the resulting resemblance metrics become invalid and misleading. Consequently, we deliberately exclude parametric measures from our proposed framework to maintain a fully non-parametric, assumption-free approach.

Abstract and method-specific metrics. Finally, some works evaluate temporal fidelity through latent space representations [18,21] or reconstruction errors specific to the generative method [29]. While latent embeddings can capture complex dependencies, they suffer from a lack of interpretability; abstract factors do not translate back to clinical measurement spaces. Furthermore, standard embedding techniques (e.g., principal component analysis) impose linearity and balance constraints, whereas non-linear alternatives further reduce transparency. Reconstruction errors, conversely, assess a method’s specific generation process rather than the continuous temporal fidelity of the underlying data. Thus, these metrics lack the generalizability required for a standardized LPD evaluation framework.

Ultimately, this assessment of both broader frameworks and individual studies reveals two major methodological gaps: the inability to handle unbalanced time steps without distortion, and an over-reliance on single-metric evaluations that can obscure structural failures. To address these limitations, we introduce a comprehensive evaluation framework tailored for unbalanced longitudinal data. Rather than relying on isolated metrics that can project a false sense of fidelity, our framework utilizes a suite of temporal preservation metrics. As Section 4 demonstrates, this combination effectively exposes the distinct ways the synthetic data may be compromised.

3. Methods

This section outlines the methodological framework of the paper: Section 3.1 introduces the preliminaries, Section 3.2 presents the proposed metrics, and Section 3.3 details the experimental setup used to illustrate them in practice.

3.1. Preliminaries

Let $i = 1, \dots, n$, index the subjects (individuals) and $j = 1, \dots, p$, index variables measured over time. For subject i and variable j , let

$$\mathbf{t}_{ij} = (t_{ij1}, \dots, t_{ijn_{ij}})', \quad n_{ij} \geq 2,$$

denote the ordered measurement times. A dataset is *unbalanced* if measurement times vary across subjects or variables, and *balanced* if they are identical. This notion of balance concerns the measurement design and differs from missing data (when a scheduled measurement is not observed) [17]. In secondary-use settings, however, missingness and structural irregularity are often indistinguishable; we therefore do not attempt to separate them.

The corresponding observations for subject i and variable j are collected into the vector $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijn_{ij}})'$ where x_{ijk} denotes the value of variable j for subject i at time t_{ijk} . Depending on the type of variable, the entries x_{ijk} may be continuous or discrete. In the latter case, they take values in a finite space $C_j = \{c_{j1}, \dots, c_{jL}\}$, which may

differ across variables. For simplicity, we refer to these sets collectively as C when the distinction is not essential.

For each variable j , subject-specific observations are combined with their measurement times t_{ij} into a data matrix $\mathbf{X}_j$ of size $\sum_i n_{ij} \times 2$. This type of representation is commonly referred to as the *long format*. An alternative is the *wide format*, in which the rows consist of transposed subject-specific vectors.

To formalize the timing of these measurements, we define the set of unique measurement times across all subjects as

$$T_j = \bigcup_{i=1}^n \{t_{ijk} : k = 1, \dots, n_{ij}\}.$$

T_j thus collects all distinct time points at which variable j was measured. To capture the presence or absence of a measurement at any given time, we define a *measurement indicator matrix*

$$\mathbf{I}_j = (I_{ijt})_{i=1, \dots, n; t \in T_j}, \quad (1)$$

with entries $I_{ijt} = \mathbf{1}\{\exists k : t_{ijk} = t\}$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. In other words, $I_{ijt} = 1$ if subject i has at least one observed measurement of variable j at time t , and $I_{ijt} = 0$ otherwise.

For each subject i , the dropout point is defined as the earliest time t after which no further measurements were made:

$$d_{ij} = \min \{t \in T_j \mid I_{ijs} = 0 \ \forall s \in T_j, s > t\}. \quad (2)$$

Throughout this work, a tilde (e.g., $\tilde{x}_{ijk}$) denotes the synthetic counterpart of the corresponding original notion.

3.1.1. Kernel smoothing

Kernel smoothing offers a flexible tool for summarizing unbalanced longitudinal data [35]. By using non-parametric smoothing, arbitrary temporal patterns can be captured, and values can be interpolated at unobserved time points. Consequently, the proposed metrics are essentially model-free, capturing universal and interpretable longitudinal behaviors.

Since the time distributions of the original and synthetic data may differ, i.e., certain time points may only appear in one of them, we evaluate all metrics for a given variable j on a common *regular* time grid:

$$T_j^* = \{t_j^{\min} + m \Delta_j : m = 0, 1, \dots, M_j\}, \quad (3)$$

where $M_j = \lfloor (t_j^{\max} - t_j^{\min}) / \Delta_j \rfloor$, with $t_j^{\min} = \min(T_j \cup \tilde{T}_j)$, $t_j^{\max} = \max(T_j \cup \tilde{T}_j)$, and $\Delta_j > 0$ is a fixed grid step (the unit of time for variable j). This grid choice defines where kernel-based estimates are evaluated, and alternative choices can be used depending on the aspects of temporal behavior one wishes to emphasize, for example by restricting evaluation to the original time range.

Definition 1 (Normalized Kernel Weights). Let $K_h(t, t')$ denote a kernel function, giving higher weights to (measurement) time points t' that are closer to $t \in T_j^*$. For example, the Gaussian kernel is defined as

$$K_h(t, t') = \frac{1}{h} \exp\left(-\frac{(t - t')^2}{2h^2}\right),$$

where $h > 0$ is the bandwidth controlling the degree of smoothing. The normalized kernel weights for the j th variable at time $t \in T_j^*$ are

$$\bar{\omega}_{ijk}(t) = \frac{K_h(t, t_{ijk})}{\sum_{i'=1}^n \sum_{k'=1}^{n_{i'j}} K_h(t, t_{i'jk'})}, \quad (4)$$

$i = 1, \dots, n, \quad k = 1, \dots, n_{ij}.$

The weights $\bar{\omega}_{ijk}(t)$ in (4) represent how much influence each observed time point t_{ijk} has when estimating or smoothing a value at the target time t . The bandwidth h determines how quickly the weights decrease, effectively setting a window of nearby points that contribute significantly to the estimate. Normalizing ensures that all weights sum

to one, allowing us to construct meaningful weighted averages at each time t .

Definition 2 introduces a weighted empirical cumulative distribution function used by some of the proposed metrics.

Definition 2 (Weighted Empirical Cumulative Distribution). The weighted empirical cumulative distribution function of the j th variable at time $t \in T_j^*$ is defined as

$$F_{jz}^{(K)}(t) = \sum_{i=1}^n \sum_{k=1}^{n_{ij}} \bar{\omega}_{ijk}(t) \mathbf{1}(x_{ijk} \leq z), \quad (5)$$

where $\bar{\omega}_{ijk}(t)$ are the normalized kernel weights (4).

Remark 1 (Choice of Bandwidth h). Kernel methods are often sensitive to the choice of bandwidth h , but in the present context this sensitivity is not a drawback. Since the aim is to assess whether real and synthetic data behave similarly, any discrepancy between the two, regardless of the value of h , can be interpreted as evidence of a difference. Thus, the bandwidth serves more as a lens that highlights differences across various smoothing scales than as a tuning parameter to be optimized. Exploring several values of h may therefore provide a more nuanced comparison. Larger values of h emphasize population-level trends, whereas smaller values highlight local fluctuations and subject-specific variation. In principle, h could be chosen systematically, see [35]. However, for simplicity, in this work, a single value of h was selected manually based on the time scale and measurement frequency of the original data.

3.1.2. Univariate temporal preservation

A central objective in this work is to evaluate how well synthetic data reproduce the temporal patterns observed in the original longitudinal data. Since variables may differ in their measurement schedules, frequencies, and temporal behaviors, it is natural to consider each variable separately.

The key idea is to summarize the temporal evolution of each variable j using a suitable metric $M_j(\cdot)$. By applying the same metric to both the original and synthetic datasets, we can assess whether the synthetic data preserves the temporal structure present in the original data.

Definition 3 (Univariate Temporal Preservation). For each variable j , consider the value–time pairs

$$\{(x_{ijk}, t_{ijk}) : i = 1, \dots, n; k = 1, \dots, n_{ij}\},$$

and let $M(\cdot)$ denote a metric that summarizes the evolution of variable j over time. Applying M_j to the original and synthetic data yields

$$M_j = M((x_{ijk}, t_{ijk})), \quad \tilde{M}_j = M((\tilde{x}_{ijk}, \tilde{t}_{ijk})).$$

We say that the synthetic data exhibit *univariate temporal preservation* for variable j if $\tilde{M}_j \approx M_j$.

Remark 2 (Goodness of Approximation). The goodness of approximation in Definition 3 is inherently data- and task-dependent. What constitutes sufficient preservation depends on the variable, the intended analyses, and the scale of variation in the data. In practice, similarity between M_j and $\tilde{M}_j$ can be assessed numerically, e.g., using tolerance thresholds or statistics such as correlation coefficients, or qualitatively through visualizations. These preservation metrics should be interpreted together with other resemblance and utility evaluations: a small discrepancy between M_j and $\tilde{M}_j$ may be inconsequential in one context but critical in another.

3.2. Metrics to assess temporal preservation

Our approach to evaluating temporal preservation in synthetic LPD is grounded in explanatory data analysis, which aims to reveal important patterns in observed data and guide the development of more formal analysis models [30]. The same principles can be used to assess how closely synthetic data resemble the original. Accordingly, we organize the key characteristics of longitudinal data into four main categories:

- I. *Marginal structure* describes population-level behavior, representing how the overall mean and proportions change over time, including potential trends and periodicity (recurring patterns).
- II. *Covariance structure* reflects how repeated measures are related over time, capturing both between-subject variability and temporal within-subject correlation.
- III. *Individual structure* focuses on subject-specific trajectories, complementing population-level summaries.
- IV. *Measurement structure* concerns the timing and frequency of measurements, including missing data, dropouts or loss to follow-up.

Table 1 presents our proposed metrics, which are further discussed in detail in the following subsections and illustrated in Section 4. All metrics can also be computed separately within appropriate subsets (strata) of the data.

3.2.1. Metrics to assess the marginal structure

Marginal summary statistics provide practical tools for evaluating temporal preservation between synthetic and original LPD. For each variable, we examine its *temporal profile*, defined as the evolution of cross-sectional distributions across the common time grid (3). For numerical variables, this includes mean and quantile trajectories; for categorical variables, class probabilities.

The *kernel-smoothed mean profile* (Metric 1) summarizes the average value at each time point by applying the normalized kernel weights (4), capturing overall trends and potential periodic patterns at the population level.

Metric 1. The kernel-smoothed mean profile of variable j at time $t \in T_j^*$ is

$$\mu_j^{(K)}(t) = \sum_{i=1}^n \sum_{k=1}^{n_{ij}} \tilde{\omega}_{ijk}(t) x_{ijk}, \quad t \in T_j^*.$$

While the smoothed mean profile effectively captures the central tendency, it does not provide information about variability or other distributional characteristics. *Kernel-smoothed quantile profiles* (Metric 2) overcome this by capturing the distribution's shape across quantiles, providing a fuller view of the variable's temporal evolution.

Mean and quantile profile plots can be further enhanced by overlaying observations outside the defined quantile range. Such visualizations are particularly useful since the mean profile and outer quantiles are sensitive to outliers, and plotting them can help explain deviations in the observed metrics and assess how well synthetic data capture outlier behavior.

Metric 2. The kernel-smoothed quantile profile of variable j at time $t \in T_j^*$ is

$$Q_{jq}^{(K)}(t) := \left\{ F_{jq}^{(K)} \right\}^{-1}(t), \quad t \in T_j^*,$$

where $F_{jq}^{(K)}$ is the weighted empirical CDF Eq. (5) and q is the quantile level of interest.

Table 1

Evaluation metrics (M) for assessing univariate temporal preservation in synthetic longitudinal patient data proposed in this work, organized by structural dimension and the corresponding aspects they diagnose.

Metric	Structural dimension	Structural deviations captured
Continuous variables		
Mean and quantile profiles (M1–2)	Marginal	Bias in mean function, distorted temporal trends, misrepresentation of distributional shape (e.g. spread, skewness, tails)
Variance profile (M4)	Covariance	Misestimation of total variance (scale) or violations of stationarity (shape)
Variogram (M5)	Covariance	Altered variance decomposition (random error, between-subject heterogeneity, autocorrelation)
Rank-order variability (M6)	Covariance	Poor preservation of relative subject ordering (e.g., different subject-level trends)
Discrete variables		
Class profiles (M3)	Marginal	Bias in class prevalence, distorted temporal composition, missing cyclic patterns
Local transition probabilities (M7)	Covariance	Misestimation of transition probabilities, state persistence, or violations of time homogeneity
Both variable types		
Subject-specific trajectories	Individual	Implausible subject-level trajectories, loss of heterogeneous progression patterns
Measurement density (M8)	Measurement	Distortion of observation time distribution (e.g. uneven sampling intensity over time)
Measurement similarity (M9)	Measurement	Discrepancy in observation patterns at the subject level (measurement schedules)
Dropout divergence (M10)	Measurement	Misrepresented dropout timing or attrition distribution

For discrete variables, the *kernel-smoothed class profile* (Metric 3) represents the proportion of each class $c_{j\ell} \in C_j$, capturing how class distributions evolve over time. These profiles can be visualized, for example, with area charts.

Metric 3. The kernel-smoothed class profile at time $t \in T_j^*$ is the L -dimensional vector $(p_{j1}^{(K)}(t), \dots, p_{jL}^{(K)}(t))$ where

$$p_{j\ell}^{(K)}(t) = \sum_{i=1}^n \sum_{k=1}^{n_{ij}} \tilde{\omega}_{ijk}(t) \mathbf{1}(x_{ijk} = c_{j\ell}), \quad \ell = 1, \dots, L.$$

3.2.2. Metrics to assess the covariance structure

Assessing the marginal structure is a fundamental first step in synthetic data evaluation and indicates whether basic distributional patterns are captured. However, for longitudinal data this is insufficient, since marginal metrics do not reflect within-subject variability

or temporal dependence. Covariance structure is therefore essential, particularly as it informs model selection in longitudinal analyses.

We start with the total variance of each variable, assessed using the *kernel-smoothed variance profile* (Metric 4), which describes how variance evolves over time. The variance profile captures both the magnitude and shape (including stationarity) of deviations from the mean profile (Metric 1).

Metric 4. The kernel-smoothed variance profile of variable j at time $t \in T_j^*$ is

$$\sigma_j^{2(K)}(t) = \sum_{i=1}^n \sum_{k=1}^{n_{ij}} \hat{w}_{ijk}(t) \left(r_{ijk}^{(K)}(t) \right)^2,$$

where the time-specific residuals are $r_{ijk}^{(K)}(t) = x_{ijk} - \mu_j^{(K)}(t)$.

In the context of LPD, repeated measurements introduce dependencies that necessitate a more detailed analysis of temporal structures, requiring tools capable of capturing autocorrelation. One such tool is the variogram (sometimes referred to as the *semivariogram*) [30,36,37], which we also estimate using kernel smoothing (Metric 5). The variogram measures variability through weighted squared differences between pairs of kernel-smoothed residuals separated by a time lag of u time units.

Metric 5. The kernel-smoothed variogram for $u > 0$ is calculated as

$$\gamma_j^{(K)}(u) = \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^{n_{ij}-1} \sum_{l=k+1}^{n_{ij}} \left(r_{ijk}^{(K)}(t_{ijl}) - r_{ijl}^{(K)}(t_{ijl}) \right)^2 \bar{w}_j^{(K)}(u, \Delta t_{ijkl}),$$

where $\Delta t_{ijkl} = t_{ijl} - t_{ijk}$ and the weights are

$$\bar{w}_j^{(K)}(u, \Delta t_{ijkl}) = \frac{K_h(u, \Delta t_{ijkl})}{\sum_{i'=1}^n \sum_{k'=1}^{n_{i'j}-1} \sum_{l'=k'+1}^{n_{i'j}} K_h(u, t_{i'jl'} - t_{i'jk'})}.$$

Assuming the total variance (Metric 4) is stationary, the variogram can be used to separate different sources of variation (Fig. 1), as it is directly related to the autocorrelation function [30]. Under stationarity, the gap between the total variance and the variogram corresponds to *between-subject variability*, the unexplained component at baseline ($t = 0$) represents *random measurement error* and the remaining part reflects temporal within-subject variability, i.e., autocorrelation.

Variogram reflects the strength of autocorrelation: a sharp rise near the origin indicates strong short-range dependence, whereas a gradual increase suggests longer-range correlation. With increasing lag, it typically approaches a plateau, corresponding to the total variance in the absence of between-subject variability. Under stationarity, values exceeding total variance generally indicate a temporal trend.

In Section 4, the kernel-smoothed variance profile and variogram are illustrated side by side as in Fig. 1. For non-stationary variables, precise analytical interpretation is secondary; the primary objective is similarity between original and synthetic metrics.

Beyond variance-based measures, it is informative to assess whether subjects preserve their relative ordering over time. To this end, we introduce a novel metric termed *kernel-smoothed rank-order variability* (Metric 6). Formally, rank-order variability $\mathcal{R}_{ij} \in [-1, 1]$ measures average directional rank drift: positive (negative) values indicate upward (downward) trends, and values near zero little systematic change.

The metrics described above for evaluating covariance structures apply to continuous variables. For example, Metric 4 has no direct analogue in the discrete setting, since the variance of a finite discrete variable is typically determined directly by its expected value.

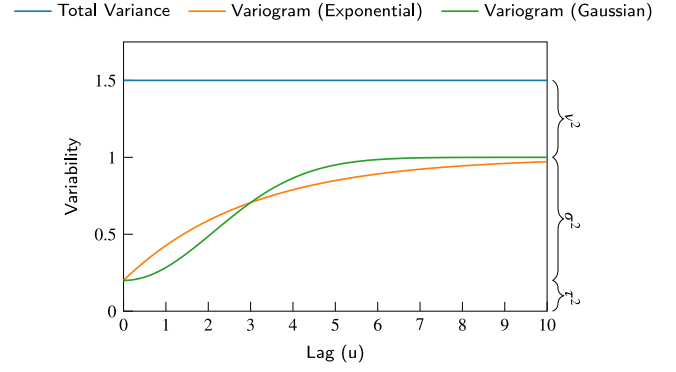


Fig. 1. Theoretical variograms under stationarity for exponential and Gaussian correlation structures. The figure illustrates how the variogram separates sources of variability: the intercept at lag zero corresponds to measurement error variance (τ^2), the increase with lag reflects within-subject serial correlation (autocorrelation) with variance component (σ^2), and the gap between the sill and the total variance represents between-subject variability due to a random intercept (v^2). Differences in curve shape indicate shorter- versus longer-range temporal dependence.

Metric 6. Let $Q = \{q_1, \dots, q_Q\}$ be a set of quantile levels, and let $Q_{jq}^{(K)}(t)$, $q \in Q$, denote the kernel-smoothed quantile profile (Metric 2) of variable j at time t . The kernel-smoothed rank-order variability for subject i and variable j is

$$\mathcal{R}_{ij} = \frac{1}{Q} \cdot \frac{1}{n_{ij} - 1} \sum_{k=1}^{n_{ij}-1} (R_{ij(k+1)}^{(K)} - R_{ijk}^{(K)}),$$

where $R_{ijk}^{(K)} \in \{1, \dots, Q+1\}$ is the quantile-bin index of x_{ijk} defined by

$$R_{ijk}^{(K)} = \ell \quad \text{if} \quad Q_{jq_{\ell-1}}^{(K)}(t) < x_{ijk} \leq Q_{jq_{\ell}}^{(K)}(t),$$

with the conventions $Q_{jq_0}^{(K)}(t) = -\infty$ and $Q_{jq_{Q+1}}^{(K)}(t) = +\infty$.

To study covariance structures in discrete variables, especially for capturing autocorrelation, we use *kernel-smoothed local transition probabilities* (Metric 7). These metrics quantify the likelihood of transitioning from one class (or state) to another over time.

Metric 7. The kernel-smoothed local transition probability from state $a \in C$ to state $b \in C$ is

$$\pi_{jab}^{(K)}(t) = \sum_{i=1}^n \sum_{k=1}^{n_{ij}-1} \bar{w}_{ijk}^{(a)}(t) \mathbf{1}(x_{ijk} = a, x_{ij(k+1)} = b),$$

where the weights are

$$\bar{w}_{ijk}^{(a)}(t) = \frac{K_h(t, t - d_{ijk}(t)) \mathbf{1}(x_{ijk} = a)}{\sum_{i'=1}^n \sum_{k'=1}^{n_{i'j}-1} K_h(t, t - d_{i'jk'}(t)) \mathbf{1}(x_{i'jk'} = a)},$$

and $d_{ijk}(t) = \max\{|t - t_{ijk}|, |t - t_{ij(k+1)}|\}$.

The resulting t -indexed collection of transition matrices $\Pi_j^{(K)}(t) = [\pi_{jab}^{(K)}(t)]_{a,b \in C}$ captures the locally smoothed probabilities. These transition probabilities reflect both *class stability*, i.e., self-transitions ($a \rightarrow a$) and change across states and are particularly useful for assessing temporal preservation in discrete variables. A variable is said to be *time-homogeneous*, analogous to stationarity in continuous settings, if its local transition probabilities remain approximately constant across time. We visualize local transitions using area charts (see Section 4.2).

3.2.3. Metrics to assess the individual structure

Many observed patterns can be better understood by visually examining *subject-specific trajectories*. For large samples, plotting all trajectories simultaneously is impractical. We therefore recommend stratification combined with subsampling. Stratification may rely on predefined groups (e.g., treatment vs. control) or, as in this study, on baseline smoothed quantile ranks. This facilitates assessing whether subjects at different points of the baseline distribution exhibit distinct temporal patterns.

3.2.4. Metrics to assess the measurement structure

In this section, we introduce a suite of novel metrics, *measurement density*, *measurement similarity*, and *dropout divergence*, specifically designed to quantify the fidelity of the measurement structure in synthetic longitudinal data.

Measurement density (Metric 8) characterizes how measurements of a given variable are distributed over time. At each time point t , it represents the proportion of all measurements of variable j occurring at t , relative to the total number of measurements across the common time grid T_j^* defined in (3). Leveraging the measurement indicators I_{ijt} (1), this common grid allows accounting for both missing observations at original time points and additional observations that appear only in the synthetic data.

Metric 8. For variable j , the measurement density at time point $t \in T_j^*$ is

$$\mathcal{MD}_{jt} = \frac{\sum_{i=1}^n I_{ijt}}{\sum_{s \in T_j^*} \sum_{i=1}^n I_{ijs}}.$$

$\mathcal{MD}_{jt}$ reveals which time points dominate the measurement schedule. A uniform density implies evenly spaced data collection, while peaks or troughs indicate periods of concentrated or scarce measurements.

To assess how closely the synthetic measurement patterns resemble the original, we use the measurement indicator matrices (1) to construct measurement similarity $\mathcal{MS}_j \in [0, 1]$ (Metric 9). A value of $\mathcal{MS}_j = 1$ indicates perfect agreement between the original and synthetic measurement patterns, while $\mathcal{MS}_j = 0$ reflects complete dissimilarity.

Metric 9. The measurement similarity of variable j is the normalized absolute distance between the original and synthetic measurement indicator matrices after optimal row alignment:

$$\mathcal{MS}_j = 1 - \frac{1}{n \cdot |T_j^*|} \sum_{i=1}^n \sum_{t \in T_j^*} |I_{ijt} - \tilde{I}_{ijt}|.$$

To assess how well synthetic data reproduce dropouts and loss to follow-up, we use *dropout divergence* (Metric 10), where a value closer to zero indicates better agreement.

Metric 10. Let $P_j(t)$ and $\tilde{P}_j(t)$ denote the empirical distributions of dropout points $\{d_{ij}\}_{i=1}^n$ (defined in Eq. (2)) in the original and synthetic data, respectively, where a small constant $\epsilon > 0$ is added to all probability masses prior to normalization to ensure positivity. The dropout divergence is the Kullback–Leibler divergence between $P_j(t)$ and $\tilde{P}_j(t)$:

$$\mathcal{D}_j(P_j \parallel \tilde{P}_j) = \sum_{t \in T_j^*} P_j \log \left(\frac{P_j(t)}{\tilde{P}_j(t)} \right).$$

Because subjects in the synthetic data do not directly correspond to those in the original data, the rows of the measurement indicator matrices must be aligned before computing Metrics 9 and 10. We obtain this alignment using the Hungarian algorithm [38], which finds the row permutation that minimizes the Euclidean distance between the rows of the real and synthetic matrices in $\mathcal{O}(n^3)$ time [39,40]. The resulting minimal distance also serves as a secondary measure of similarity between the two matrices.

This alignment requires measurement matrices of identical dimensions. When the numbers of synthetic and original subjects differ, or when the Hungarian algorithm becomes computationally expensive, equal-sized random subsamples can be drawn from both matrices. To limit variability due to random subsampling, the procedure is repeated multiple times and the resulting scores averaged.

Interpreting the resulting scores requires a reference. One option is to use a separate test set or, as described in Section 3.3, a resampled subset of the original data. This enables comparison between original-to-synthetic and original-to-original samples, with the former ideally matching the latter.

3.3. Experimental design

To illustrate the proposed metrics, we generated synthetic longitudinal patient data using two SDG methods: HALO [23] and Health Gym GAN (HGG) [33]. HALO is GPT-2-based and generates unbalanced data directly, whereas HGG is GAN-based and requires prior imputation with missingness indicators, with unbalancedness reintroduced after generation. Both methods support mixed continuous and categorical variables and are publicly available.

We followed the original codebases throughout the generation process, including dataset selection (MIMIC-III [41,42]), cohort definitions (HALO: 32,060 subjects, 17 variables; HGG: 3910 subjects, 13 clinical variables plus 7 missingness indicators), and preprocessing. The examples are intended to demonstrate the proposed metrics rather than to enable direct method comparison.

To harmonize illustrations, HALO outputs were post-processed by converting time intervals to a continuous hourly scale and restricting analyses to the first 48 h, motivated by measurement sparsity beyond this window. The HGG acute hypotension subset [43] is inherently limited to 48 h. Bandwidth $h = 6$ was used throughout.

Measurement similarity and drop-out divergence (Metrics 9–10), together with Frobenius norms of the measurement indicator matrices (1), were computed on random samples of 2000 subjects and repeated over 100 iterations due to the computational complexity of the Hungarian algorithm. Reported values represent iteration averages, with reference values (in parentheses) obtained from separate samples of the original data.

4. Results

This section illustrates the evaluation metrics introduced in Section 3.2, with emphasis on graphical methods. Given the complexity and heterogeneity of LPD, visualizations complement numerical metrics by facilitating intuitive assessment of temporal patterns, trends, and anomalies that may not be apparent from scalar summaries alone.

For pedagogical clarity, we follow the order of the four main longitudinal data characteristics described in Section 3.2. However, these metrics can be applied in any order (see further discussion in Section 6).

Table 2 summarizes the selected examples, chosen to highlight distinct deviations in temporal preservation (Appendix A includes an example with less evident deviations). The following subsections discuss each case in detail, with additional figures included in Appendix B and referenced as supplementary figures (e.g., Fig. S.1, Fig. S.2).

Table 2
Overview of illustrative examples* used to demonstrate how the proposed evaluation metrics diagnose specific deviations in temporal and distributional structure.

Variable	Structural dimension affected	Observed deviation	Indicative metrics
Weight	Covariance structure	Loss of between-subject heterogeneity despite preserved marginal distribution	Variogram, rank-order variability
Respiratory rate	Covariance structure	Distorted variability and inflated extreme values with partially preserved autocorrelation	Variance profile, variogram, marginal profiles with outliers
Aspartate aminotransferase	Marginal and covariance structure	Suppression of distributional tails, reduced variance and opposite subject-level trends	Marginal profiles with outliers, variance profile, variogram, rank-order variability
GCS score	Transition probability structure	Unstable or inconsistent transition dynamics in minority classes	Local transition probabilities
GCS Eye opening assessment	Marginal structure	Spurious class introduction in synthetic distribution	Class profiles
FiO ₂	Marginal structure	Missing class in synthetic distribution	Class profiles

* The analyses were conducted for all variables; the examples highlight representative phenomena across different structural deviations, while omitted variables exhibited comparable patterns.

4.1. Problems in preserving the covariance structure

Failures in covariance structure preservation are illustrated using weight and respiratory rate generated by HALO and aspartate aminotransferase (AST) generated by HGG. These examples demonstrate that closely matching marginal behavior and numerical scores can still coincide with substantial differences in covariance structure in the synthetic data.

Changes in variability decomposition. Fig. 2(a) depicts the mean and quantile profiles (Metrics 1 and 2) of the variable weight, which appear remarkably similar between the original and synthetic datasets. As data sparsity can influence synthetic data generation, the bottom panels of Fig. 2(a) also illustrate the proportions of patients under follow-up over time.

The variance profile (Metric 4) and variogram (Metric 5) in Fig. 2(b) reveal critical differences. While the variance profiles largely agree, the variograms differ substantially. The low and flat variogram in the left panel appropriately captures the fact that individual weight remains stable during the 48-hour follow-up, with almost all variability following from between-subject variability (see Fig. 1 for guidance on interpretation). In contrast, the synthetic data demonstrate a markedly different decomposition of variability, with reduced between-subject heterogeneity and a large contribution from measurement error. This change in covariance structure is also evident in the failure to reproduce rank-order variability (Metric 6; Fig. 3(a)).

The subject-specific trajectories depicted in Fig. 3(b) provide further insight into the differences in covariance structure. In the synthetic data, measurements exhibit pronounced short-term fluctuations and appear largely as random noise, with little evidence of consistent temporal patterns within individuals. The synthetic data also contain a larger number of subjects with only a single observation, visible as isolated points in Fig. 3(b). Importantly, these differences are not driven by outliers (Fig. S.6).

The measurement density closely (Metric 8) matches the original distribution (Fig. S.7), although the synthetic data exhibit a propensity for increased measurements during the initial follow-up period. The mean Frobenius norm (smaller is better) between the original and synthetic measurement matrices was 35.21 (reference 30.63), the mean measurement similarity score (Metric 9, larger is better) was 0.99 (reference 0.99), and the mean drop-out divergence (Metric 10, smaller is better) was 0.14 (reference 0.02). Overall, aside from a slightly higher dropout divergence, these numerical scores closely match the reference values.

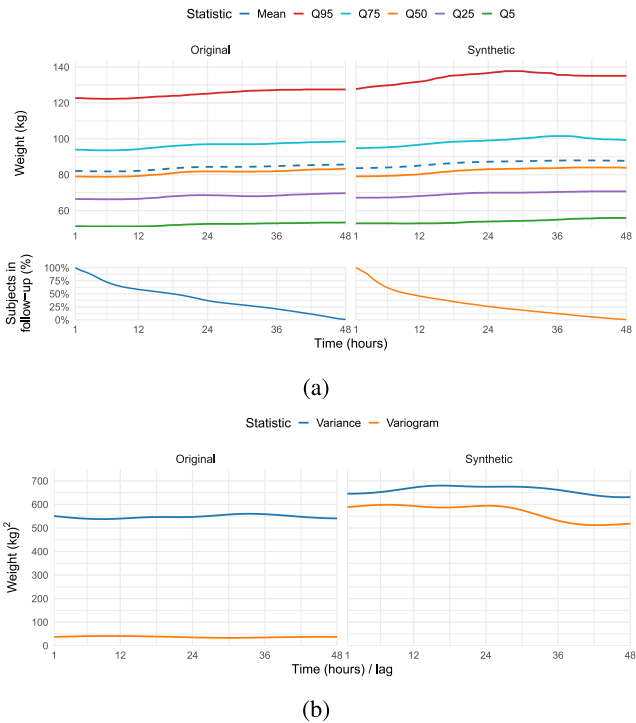


Fig. 2. Evaluation of weight (original vs. synthetic). (a) Mean and quantile profiles (Metrics 1 and 2) show close agreement between the original and synthetic data, while the lower panels indicate the proportion of patients under follow-up over time. (b) Variance profiles and variograms (Metrics 4 and 5) reveal important differences in variability decomposition: although overall variance is similar, the synthetic data exhibit reduced between-subject heterogeneity and increased measurement-error variability compared with the original data.

Inflated variance due to outliers. Mean and quantile profiles for respiratory rate (Fig. S.8) illustrate a predominantly preserved quantile ordering, with a marginally elevated mean and a lower 5th percentile compared to the original data. However, the variance profile in the synthetic data is substantially inflated (Fig. 4(a)). Yet, the rank-order variability (Fig. 4(b)) indicates that individual rankings are largely preserved, i.e., autocorrelation structure is reasonably maintained. Indeed, the shape of the variogram relative to the variance scale is adequately preserved in the synthetic data (Fig. S.9).

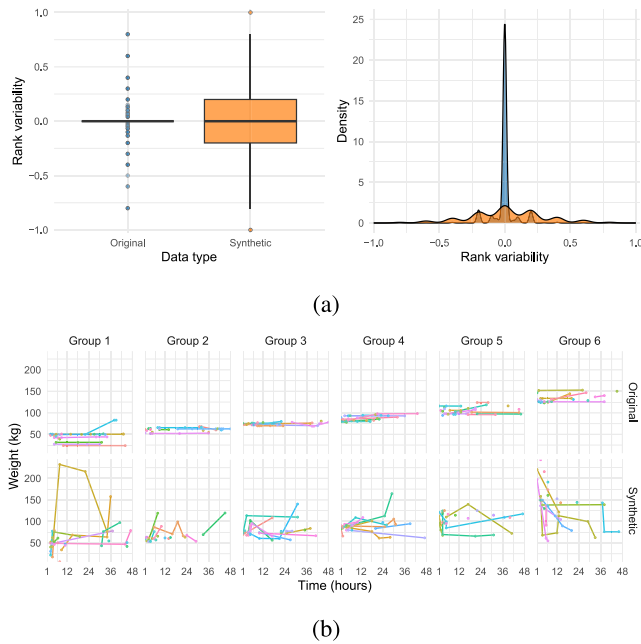


Fig. 3. Evaluation of weight (original vs. synthetic). (a) Distribution of rank-order variability (Metric 6), showing that the synthetic data do not adequately reproduce the rank stability observed in the original data. (b) Subject-specific trajectories illustrating further covariance-structure differences: synthetic trajectories show pronounced short-term fluctuations, weaker within-subject temporal consistency, and a larger number of subjects with only a single observation.

An examination of subject-specific trajectories (Fig. S.10) reveals that while some synthetic subjects follow patterns resembling those in the original data, many exhibit abrupt or erratic changes. This supports the earlier conclusion that the SDG method has preserved some degree of autocorrelation.

When the mean-quantile profiles are supplemented by plotting observations outside the defined quantile range (Fig. 4(c)), we see that the synthetic data contain outliers across a much wider range. This helps explain the exaggerated variance profile and abrupt changes in the subject-specific trajectories. Moreover, the distribution of these outliers appears almost uniformly random, in contrast to the behavior seen in the original data.

The mean Frobenius norm between the original and synthetic measurement matrices was 166.87 (reference 61.32), the mean measurement similarity score was 0.71 (reference 0.96), and the mean drop-out divergence was 0.28 (reference 0.03). These indicate non-negligible deviations in the measurement structure relative to the reference.

Deflated variance due to data sparsity. Mean and quantile profiles for aspartate aminotransferase (Fig. S.11) demonstrate a markedly attenuated 95th percentile and fail to replicate the increasing trend observed in the original data. The variance profile is substantially deflated (Fig. 5(b); see also Fig. S.12 for an unconstrained y-axis scale). In the original data, the variogram increases rapidly with lag and exceeds the total variance, indicating a temporal trend that is not reflected within the synthetic data. Moreover, rank-order variability displays opposite behavior (Fig. 5(c)): values decrease over time in the original data but increase in the synthetic data.

The subject-specific trajectories (Fig. S.13) confirm greater variability among individuals above the 95th percentile (group 6) in the original data. Although trajectories are generally stable over time (median around zero), individuals in this group often exhibit decreasing trends, consistent with the rank-order variability. This pattern is largely absent in the synthetic data.

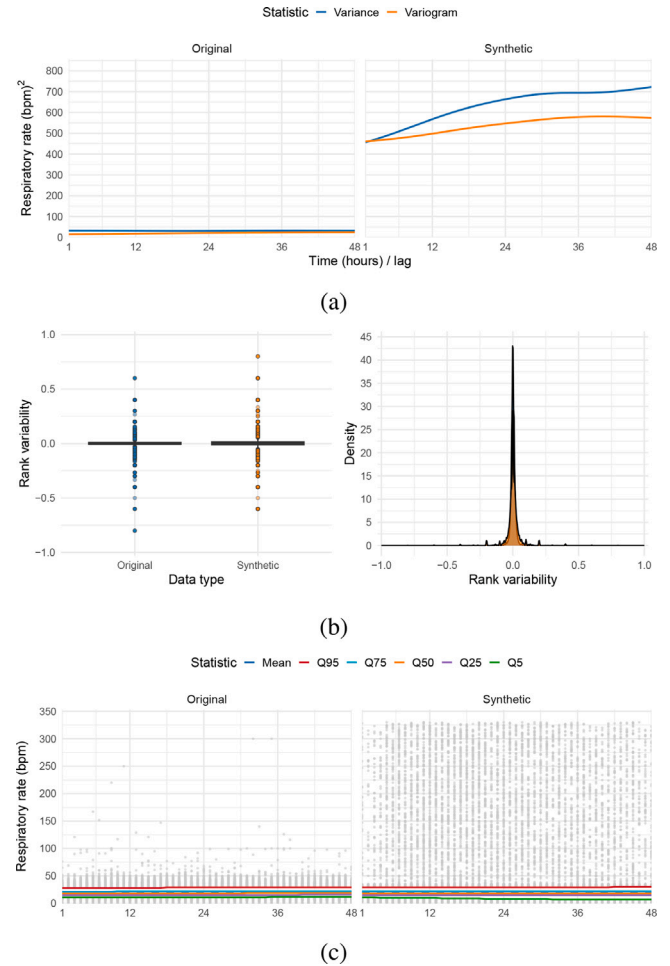


Fig. 4. Evaluation of respiratory rate (original vs. synthetic). (a) Variance profile and variogram, showing that variability is substantially inflated in the synthetic data despite preservation of a similar shape of the temporal dependence structure (with unconstrained y-axis scale provided in Fig. S.9); (b) rank-order variability, indicating that individual rankings are largely preserved and that the autocorrelation structure is reasonably maintained; (c) mean and quantile profiles, including observations outside the main quantile range, indicating that the inflated variance is driven by pronounced outliers in the synthetic data.

The mean-quantile plot with outliers (Fig. 5(a)) further reveals that the synthetic data contain fewer extreme values than the original data, in contrast to the behavior observed for the respiratory rate (Fig. 4(c)). Together with the suppressed 95th percentile, this reduction in outliers explains the deflation of total variance.

The synthetic data reproduced similar measurement density (Fig. S.14). The mean Frobenius norm between the original and synthetic measurement matrices was 42.66 (reference 24.13), the mean measurement similarity score was 0.98 (reference 0.99), and the mean drop-out divergence was 0.07 (reference 0.01), suggesting minor deviations in terms of measurement structure.

4.2. Discrepancies between class proportions and transition probabilities

Finally, we illustrate the proposed metrics for discrete variables. The class profiles (Metric 3) of Glasgow Coma Scale (GCS) score generated by HGG (Fig. 6(a)) are largely preserved, even with 13 distinct classes. Deviations are more pronounced for less frequent classes in the synthetic data.

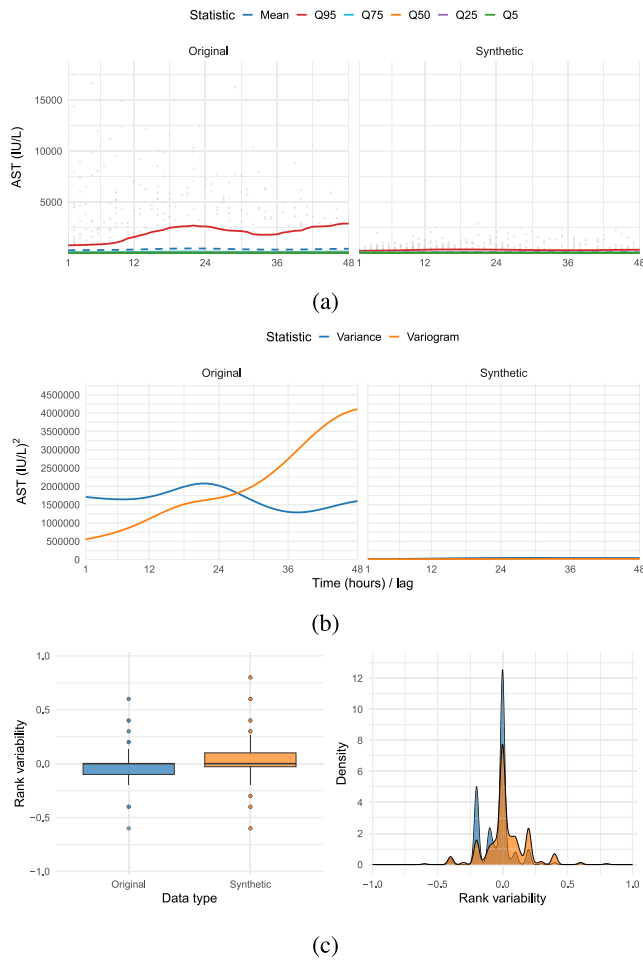


Fig. 5. Evaluation of aspartate aminotransferase (original vs. synthetic). (a) Mean and quantile profiles, including observations outside the quantile range, revealing a markedly attenuated 95th percentile and failure to reproduce the increasing trend observed in the original data. (b) Variance profile and variogram, showing substantially deflated variance in the synthetic data; an unconstrained y-axis scale is provided in Fig. S.12. (c) Rank-order variability, indicating a reversal of temporal ranking behavior, where values decrease over time in the original data but increase in the synthetic data.

Transition probabilities (Metric 7) are reasonably well captured for the majority class “R-15” (Fig. 6(b)), whereas those of minority classes are not learned reliably (Fig. 6(c)). This behavior was consistent across discrete variables with both SDG methods (see additional examples in Figs. S.15 and S.16).

The mean Frobenius norm of GCS score between the original and synthetic measurement matrices was 121.64 (reference 62.82), the mean measurement similarity score was 0.85 (reference 0.96), and the mean drop-out divergence was 0.12 (reference 0.02). These results indicate a noticeable discrepancy in the measurement structure.

In general, discrepancies were most evident in the treatment of minority classes. The GCS Eye Opening assessment generated by HALO included an additional class, “none,” absent from the training data (Fig. S.17(a)), whereas the FiO₂ variable generated by HGG failed to reproduce the class “Range-0.0” (Fig. S.17(b)). The results illustrate that synthetic data can exhibit discrepancies in both class proportions and transition probabilities, with rare classes being particularly difficult to reproduce reliably.

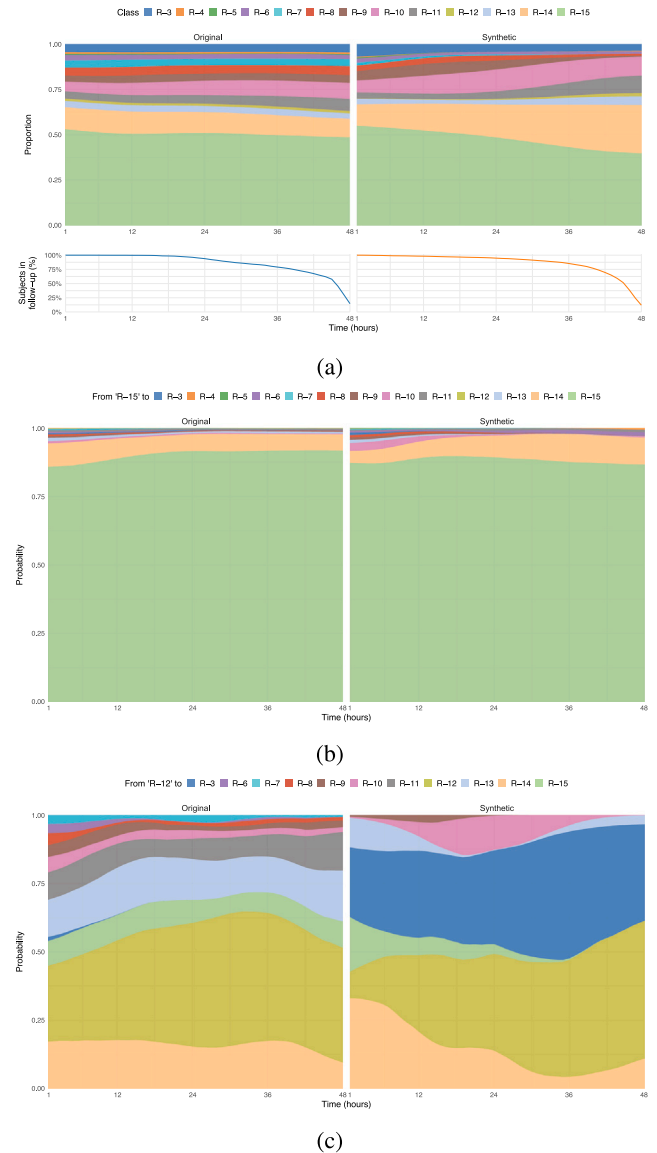


Fig. 6. Evaluation of Glasgow Coma Scale (GCS) score (original vs. synthetic). (a) Class profiles (Metric 3), showing that the overall distribution across 13 classes is largely preserved, with larger deviations observed for less frequent classes. (b) Local transition probabilities from class R-15 (Metric 7), which are well captured for the majority class. (c) Local transition probabilities from class R-12, illustrating substantially weaker agreement for minority classes.

5. Discussion

We introduced a comprehensive set of metrics for evaluating temporal preservation in synthetic longitudinal patient data. Recent advances in generative modeling have enabled the creation of increasingly complex synthetic longitudinal datasets. However, evaluation frameworks have largely remained focused on cross-sectional similarity, neglecting temporal dependencies essential to LPD. Our proposed metrics address this gap by providing tools to quantify how well synthetic data reproduce key longitudinal characteristics, including marginal, covariance and individual structures as well as measurement patterns.

Our illustrations reveal that marginal agreement does not guarantee temporal fidelity. Although weight aligned well in terms of means and quantiles, variogram analysis and subject-specific trajectories revealed distorted covariance and irregular individual patterns. Marginal similarity alone can therefore be misleading. Joint evaluation

across complementary metrics provides a more reliable assessment of temporal preservation and clarifies where structural deficiencies arise.

Across metrics, several factors shaped temporal preservation. First, synthetic quality depended on the original data and their preprocessing: outliers and implausible values (e.g., negative measurement times) were reproduced with altered magnitude, distorting covariance structure. This supports systematic cleaning unless extreme values are substantively relevant. In most longitudinal applications, however, the primary goal is to model overall disease behavior or predict patient outcomes, rendering such outliers rather noise than signal.

Second, representation choices mattered. Discretization of continuous variables, followed by reconstruction via within-bin sampling, may introduce artifacts when bin widths are large or class granularity is high. In our results, an excessively wide bin produced synthetic outliers (Fig. 4(c)). Severe imbalance may also induce mode collapse [34], where minority classes are not generated (Fig. S.17(b)) and transition probabilities become unreliable (Fig. 6(c)). Data-driven discretization, such as density-based binning [44], may mitigate these effects.

Measurement density was another key determinant. Frequently measured variables (e.g., systolic blood pressure; Fig. S.1) were reproduced more reliably than sparsely observed ones, such as aspartate aminotransferase (Fig. 5(a)), recorded at only 1.5% of time points. Uneven measurement schedules thus lead to variable-specific discrepancies, highlighting the need for SDG methods that explicitly accommodate unbalanced LPD.

Temporal encoding further affected fidelity. In our setting, the empirical data were strongly concentrated at full-hour intervals (99.7%), whereas the synthetic data showed a substantially lower proportion of such alignments (13.1%), resulting in overlapping or alternating time points after rescaling and characteristic zigzag patterns in the measurement density profiles (e.g., Fig. S.5). This illustrates that rescaling or mismatch between the empirical observation times and the representation used for SDG can distort measurement-time distributions, underscoring the importance of consistent temporal encoding.

Finally, model-specific assumptions introduced additional challenges. Fixed vocabularies in language model-based generators can produce artifacts when rare classes are unseen during training yet remain generatable (Fig. S.17(a)), motivating dynamic vocabulary construction [45]. GAN-based approaches are sensitive to missing-data handling: forward-fill or mean imputation combined with sparsity can yield overly smooth trajectories with reduced variability, as observed for aspartate aminotransferase (Fig. S.11); more advanced imputation strategies may better preserve variability [46].

5.1. Limitations

This study has several limitations. First, our objective was not to compare specific SDG methods but to illustrate the proposed resemblance metrics. A fair benchmarking study was infeasible because different SDG approaches require distinct preprocessing pipelines, precluding a common training set. Moreover, the continuous-variable version of HALO [23] is only briefly described in the original publication and may still be under development, which could explain some of the preprocessing choices. For both HALO and HGG, we trained a single generator; comprehensive evaluation would require multiple runs across datasets and hyperparameter settings, which was beyond the scope of this work.

Second, we did not assess logical temporal constraints such as monotonicity (values expected to change only in one direction over time) or non-recurrence (values that, once observed, should not revert, e.g., a specific gene mutation). These constraints were not meaningfully represented in our data but are important features of realistic longitudinal processes and should be incorporated in future evaluations.

Third, our analysis focused on univariate resemblance metrics. While multivariate temporal comparisons are relevant, they are often

difficult to interpret and visualize. Moreover, substantial univariate deviations already signal structural deficiencies. Developing interpretable multivariate temporal preservation metrics remains an important direction for future research.

Fourth, we did not explicitly distinguish between true missingness and structural unbalancedness. Established diagnostics for missing data, such as visualization of missingness patterns, Little's MCAR test [47, 48], or predictive models of missingness, could help disentangle these mechanisms. Adapting such approaches to real-world longitudinal patient data represents a promising avenue for further work.

Finally, we did not evaluate privacy protection or downstream analytical utility. Structural resemblance metrics do not directly quantify re-identification risk or task-specific performance, and their relationship to privacy and utility is not straightforward [49–51]. Future work should therefore complement temporal resemblance assessment with dedicated privacy and utility evaluations to better characterize these trade-offs.

6. Conclusion

We introduced a set of resemblance metrics for evaluating univariate temporal preservation in synthetic longitudinal patient data, addressing the limitations of evaluation frameworks that focus primarily on cross-sectional similarity. These metrics characterize four fundamental structural dimensions: marginal (population-level distributions), covariance (total variability, between-subject heterogeneity, and autocorrelation), individual (subject-specific trajectories), and measurement (observation frequency and timing).

Our examples demonstrate that marginal agreement can mask substantial distortions in covariance structure, individual structure, and measurement patterns. Temporal fidelity was strongly shaped by data quality, preprocessing, and modeling assumptions, with sparse observations, class imbalance, and mismatches in time representation emerging as recurring challenges. Overall, the proposed metrics offer a structured and interpretable framework to diagnose where temporal structure is preserved or degraded.

From a practitioner's perspective, subject-specific trajectories offer an intuitive visual diagnostic of whether synthetic data plausibly reproduce temporal behavior—ranging from manifest discrepancies to subtle, systematic deviations. However, the remaining metric categories are indispensable for identifying the structural etiology of anomalies, particularly those that remain visually imperceptible within the trajectories themselves.

Rigorous evaluation across all four characteristics is essential for the responsible use of synthetic longitudinal data. Moreover, the proposed framework is designed to complement, rather than replace, assessments of privacy and utility. Future work should extend the proposed metrics to multivariate settings and incorporate considerations for privacy protection and analytical utility.

CRedit authorship contribution statement

Katariina Perkonoja: Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Parisa Movahedi:** Writing – original draft, Visualization, Formal analysis, Data curation, Conceptualization. **Antti Airola:** Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Kari Auranen:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Joni Virta:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Use of generative AI

In the preparation of this work, the authors used ChatGPT (OpenAI, Version GPT-5) to assist in improving the English language and conciseness of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the final manuscript.

Funding

This work has received funding from European Union's Horizon Europe research and innovation programme (grant number 101095384). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. The work of KP, JV and AA was supported by the Research Council of Finland (grants 347501, 353769, 358868, 368494). KP also received support from the Finnish Cultural Foundation (grant 00260122). The funding agencies had no involvement in the study.

Declaration of competing interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and / or publication of this article.

Acknowledgments

We gratefully acknowledge Auria Clinical Informatics and Arho Virkki from Turku University Hospital for providing the computational environment that enabled this research.

Appendix A. Moderate temporal preservation

Systolic blood pressure generated by HALO demonstrates moderate temporal preservation according to the proposed evaluation metrics. Fig. S.1 illustrates that the kernel-smoothed mean profile (Metric 1) and quantile profiles (Metric 2) of synthetic data closely match those of the original data. The main deviation occurs at the 95th percentile, which is higher in the synthetic data, indicating slightly deviating outlier patterns. No clear trends or periodic patterns emerge. The original data demonstrate a pronounced reduction in the number of patients just before 48 h, whereas the synthetic data exhibit a more gradual decline.

The relatively stable variance profile (Metric 4, Fig. S.2, upper curve) suggests a stationary process that is reasonably preserved in the synthetic data. The variogram (Metric 5, Fig. S.2, lower curve) further indicates gradually declining autocorrelation and, because it does not reach the variance profile, the presence of between-subject variability, both of which are well preserved in the synthetic data.

To further examine the covariance structure of systolic blood pressure, we analyzed the kernel-smoothed rank-order variability (Metric 6, Fig. S.3). In both datasets, profiles are centered near zero, indicating that individuals generally maintain stable rankings relative to baseline, consistent with the autocorrelation observed in the variogram.

Subject-level trajectories (Section 3.2.3) can offer deeper insights into population-level metrics. The raw unsmoothed trajectories (Fig. S.4), presented in six baseline quantile-defined percentile classes [0, 5], (5, 25], (25, 50], (50, 75], (75, 95], (95, 100], resemble the original data overall, with slightly more variation in groups 1 and 6. These patterns are directly connected to the increased variance in Fig. S.2 and the elevated 95th percentile in Fig. S.1.

Fig. S.5 depicts the measurement density (Metric 8) for both datasets. In the synthetic data, observations are more frequent at earlier time points but become more sparse after about 18 h, whereas in the original data, measurements increase rapidly up to 6 h and then gradually decrease, with a sharp drop after 42 h. The synthetic series also displays strong fluctuations between adjacent time points, forming a zigzag pattern that suggests a periodic artifact of the data generating mechanism.

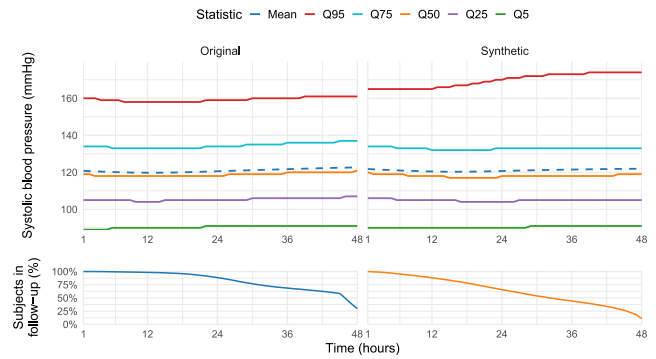


Fig. S.1. Kernel-smoothed mean (Metric 1) and quantile profiles (Metric 2) of systolic blood pressure for the original data (left) and synthetic data (right).

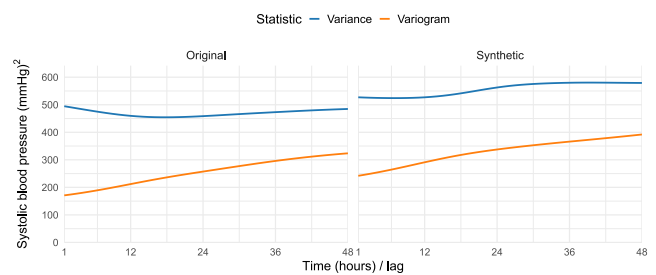


Fig. S.2. Kernel-smoothed variance profile (Metric 4, blue) and variogram (Metric 5, orange) of systolic blood pressure for the original data (left) and synthetic data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

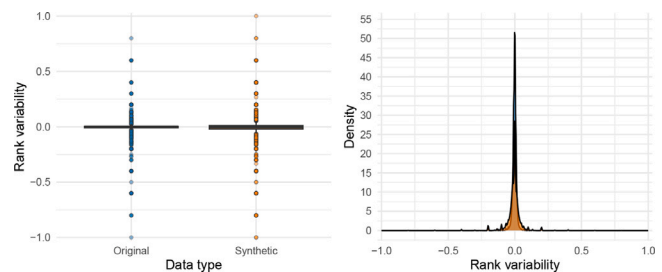


Fig. S.3. Boxplot (left) and density plot (right) illustrating the kernel-smoothed rank-order variability (Metric 6) of systolic blood pressure for original (blue) and synthetic data (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

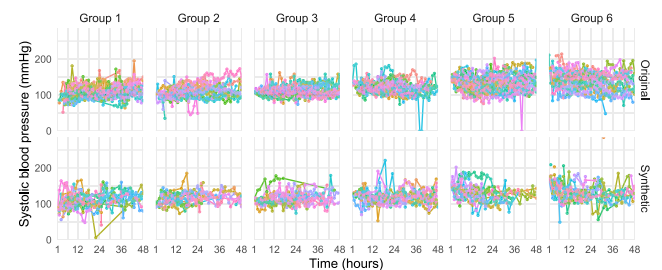


Fig. S.4. Subject-specific systolic blood pressure trajectories by baseline group for the original (top) and synthetic data (bottom).

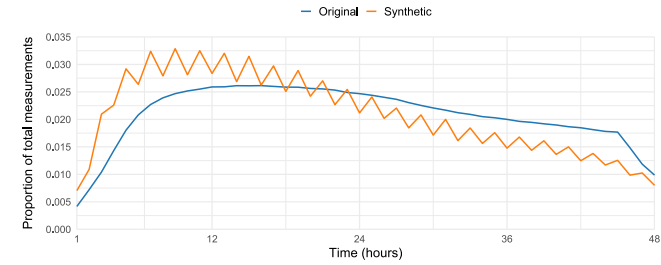


Fig. S.5. Measurement density (Metric 8) of systolic blood pressure for the original data (blue) and synthetic data (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The mean Frobenius norm (smaller is better) between the original and synthetic measurement matrices was 173.42 (reference 67.09). The mean measurement similarity score (Metric 9, larger is better) was 0.69 (reference 0.95), and the mean drop-out divergence (Metric 10, smaller is better) was 0.31 (reference 0.03). Compared to the natural variation within the real data, the synthetic data exhibit substantially greater differences in measurement structure, as evidenced by the higher Frobenius norm and drop-out divergence. Nevertheless, the measurement similarity remains comparatively high, consistent with the patterns identified across the other metrics.

Appendix B. Figures

See Figs. S.6–S.17.

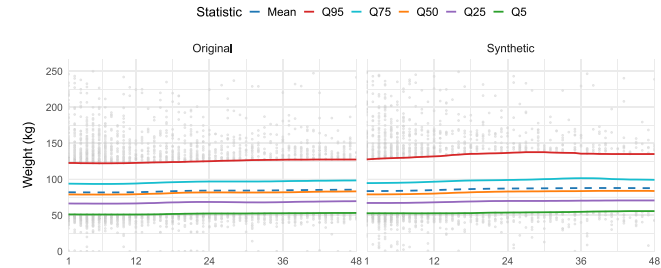


Fig. S.6. The mean–quantile profiles of weight same as in Fig. 2(a) augmented with observations below the 5th percentile and above the 95th percentile.

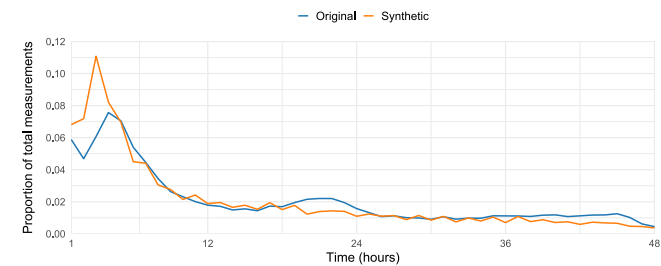


Fig. S.7. Measurement density of weight for the original data and synthetic data.

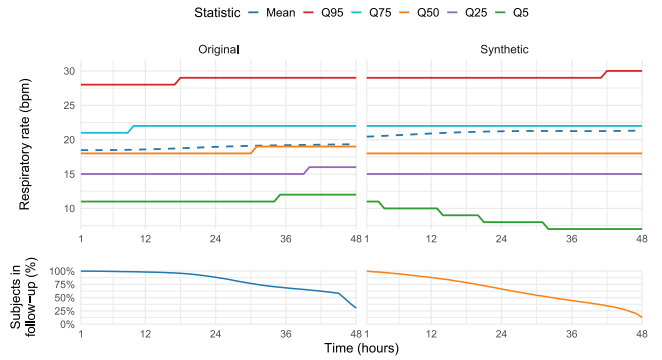


Fig. S.8. Kernel-smoothed mean and quantile profiles of respiratory rate for the original data and synthetic data.

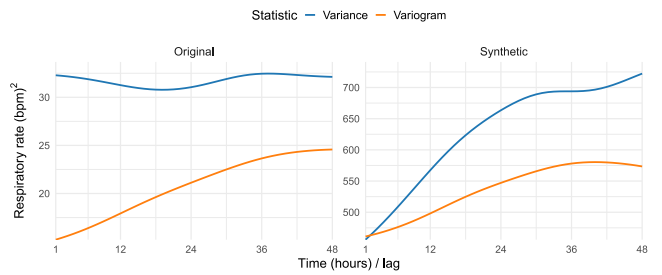


Fig. S.9. The figure replicates Fig. 4(a) with separate y-axes. While the variogram exhibits an almost identical shape across the two datasets, the variance profile differs in both scale and form compared with the original data.

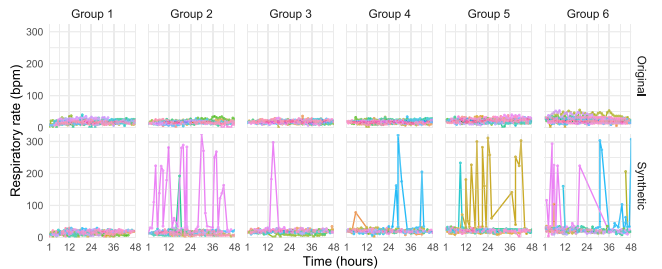


Fig. S.10. Subject-specific respiratory rate trajectories.

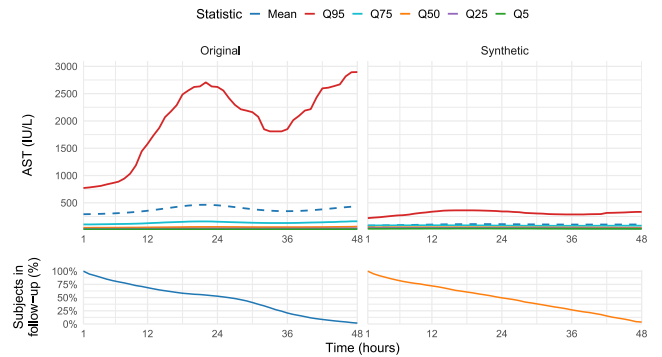


Fig. S.11. Kernel-smoothed mean and quantile profiles of aspartate amino-transferase for the original data and synthetic data.

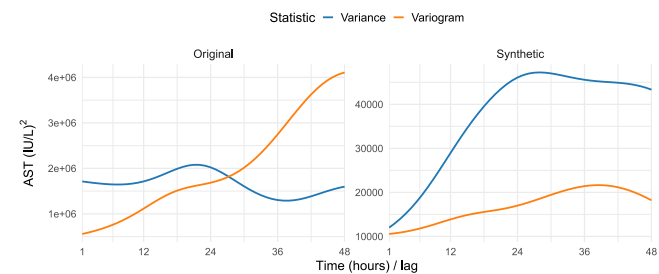


Fig. S.12. The figure replicates Fig. 5(b) with separate y-axes.

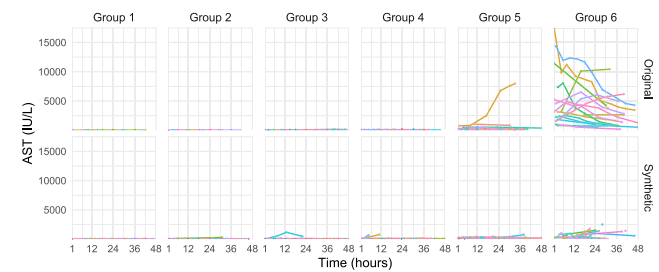


Fig. S.13. Subject-specific aspartate aminotransferase trajectories.

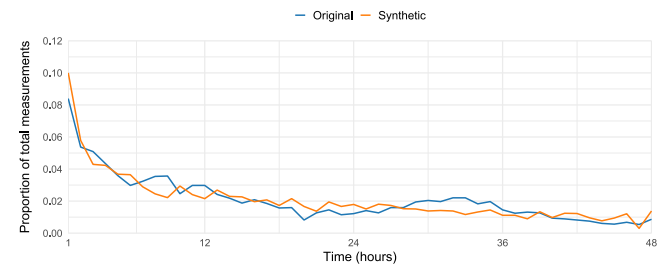


Fig. S.14. Measurement density of aspartate aminotransferase.

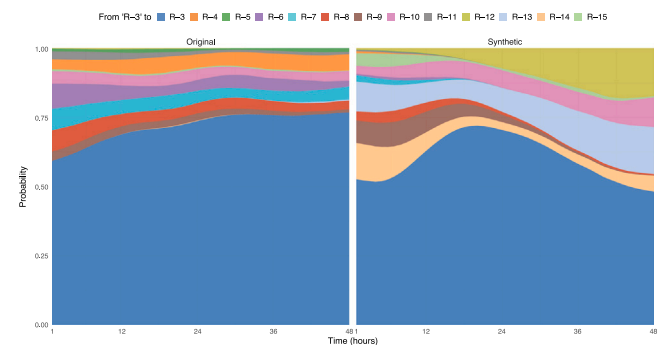


Fig. S.15. Transition probabilities of class R-3 of Glasgow Coma Scale score.

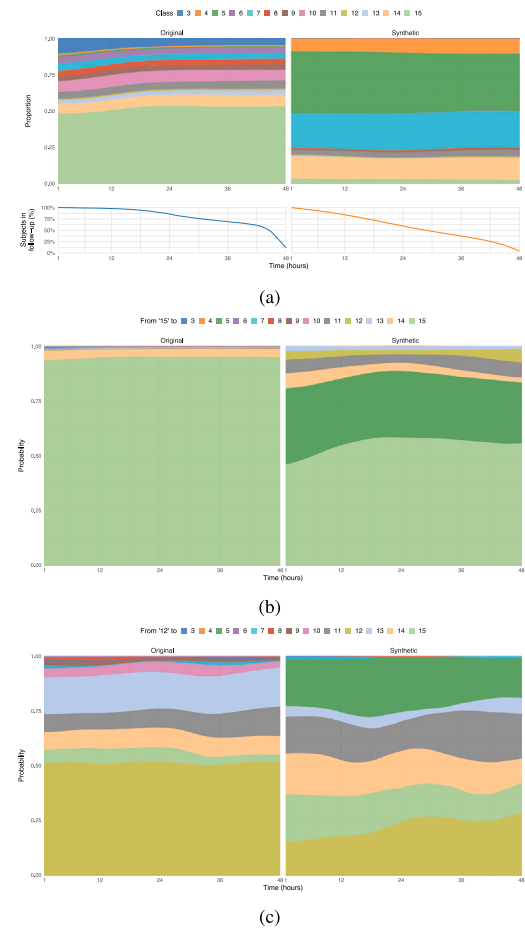


Fig. S.16. Glasgow Coma Scale scores generated by HALO, depicting (a) the kernel-smoothed class profiles, (b) transition probabilities originating from class R-15, (c) transition probabilities originating from class R-12.

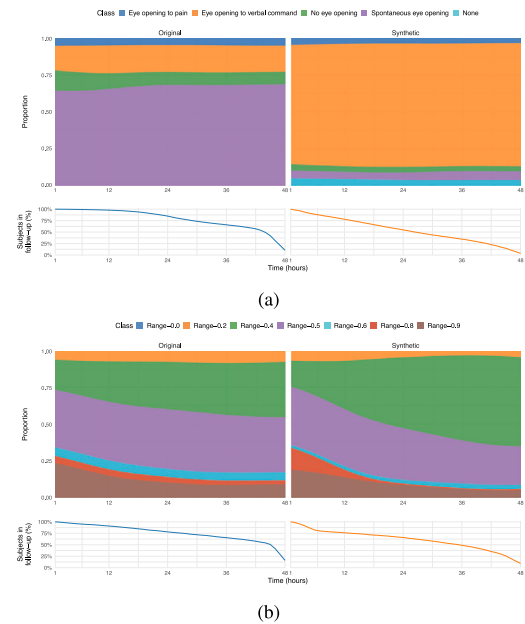


Fig. S.17. Kernel-smoothed class profiles illustrated for the original data and the synthetic data. Shown here is (a) the Glasgow Coma Scale Eye Opening assessment generated by HALO, and (b) FiO₂ generated by HGG.

Data availability

The data used in this study are derived from the Medical Information Mart for Intensive Care (MIMIC-III) database, which is publicly available through the PhysioNet repository after completion of the required credentialed data use process.

The data-extraction scripts and Health Gym GAN source code are available at https://github.com/Nic5472K/ScientificData2021_HealthGym, while the HALO implementation is provided at https://github.com/btheodorou99/HALO_Inpatient.

The source code accompanying this work, available at <https://github.com/Katsk1/Temporal-preservation-metrics>, provides the implementation of the proposed evaluation metrics and the analysis pipeline used in this study.

References

- [1] K. Perkonoja, K. Auranen, J. Virta, Methods for generating and evaluating synthetic longitudinal patient data: A systematic review, *J. Heal. Informatics Res.* 10 (1) (2026) 1–39, <http://dx.doi.org/10.1007/s41666-025-00223-7>.
- [2] G.O. Ghoshheh, J. Li, T. Zhu, A survey of generative adversarial networks for synthesizing structured electronic health records, *ACM Comput. Surv.* 56 (6) (2024) 1–34, <http://dx.doi.org/10.1145/3636424>.
- [3] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, D. Rankin, Synthetic data generation for tabular health records: A systematic review, *Neurocomputing* 493 (2022) 28–45, <http://dx.doi.org/10.1016/j.neucom.2022.04.053>.
- [4] A.D. Lautrup, T. Hyrup, A. Zimek, P. Schneider-Kamp, Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data, *ACM Comput. Surv.* (2024) <http://dx.doi.org/10.1145/3704437>.
- [5] H. Murtaza, M. Ahmed, N.F. Khan, G. Murtaza, S. Zafar, A. Bano, Synthetic data generation: State of the art in health care domain, *Comput. Sci. Rev.* 48 (2023) 100546, <http://dx.doi.org/10.1016/j.cosrev.2023.100546>.
- [6] M. Ibrahim, Y.A. Khalil, S. Amirajab, C. Sun, M. Breuwer, J. Pluim, B. Elen, G. Ertaylan, M. Dumontier, Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges, *Comput. Biol. Med.* 189 (2025) 109834, <http://dx.doi.org/10.1016/j.combiomed.2025.109834>.
- [7] Towards European Health Data Space (TEHDAS), Draft Guideline on Data Minimisation, Pseudonymisation, Anonymisation and Synthetic Data, Technical Report, 2025.
- [8] A.M. Alaa, B. Van Breugel, E. Saveliev, M. Van Der Schaar, How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Baltimore, Maryland, 2022, pp. 290–306.
- [9] K. Bhanot, M. Qi, J.S. Erickson, I. Guyon, K.P. Bennett, The problem of fairness in synthetic healthcare data, *Entropy* 23 (9) (2021) 1165, <http://dx.doi.org/10.3390/e23091165>.
- [10] M. Pereira, M. Kshirsagar, S. Mukherjee, R. Dodhia, J. Lavista Ferres, R. de Sousa, Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data, in: T.-D. Cao (Ed.), *PLoS One* 19 (2) (2024) e0297271, <http://dx.doi.org/10.1371/journal.pone.0297271>.
- [11] P. Diggle, P. Heagerty, K.-Y. Liang, S.L. Zeger, Introduction, in: *Analysis of Longitudinal Data*, second ed., Oxford University Press, 2002, pp. 1–21, <http://dx.doi.org/10.1093/oso/9780198524847.001.0001>.
- [12] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, D. Rankin, Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions, *Methods Inf. Med.* 62 (S 01) (2023) e19–e38, <http://dx.doi.org/10.1055/s-0042-1760247>.
- [13] F.K. Dankar, M.K. Ibrahim, L. Ismail, A multi-dimensional evaluation of synthetic data generators, *IEEE Access* 10 (2022) 11147–11158, <http://dx.doi.org/10.1109/ACCESS.2022.3144765>.
- [14] C. Yan, Y. Yan, Z. Wan, Z. Zhang, L. Omberg, J. Guinney, S.D. Mooney, B.A. Malin, A multifaceted benchmarking of synthetic electronic health record generation models, *Nat. Commun.* 13 (1) (2022) 7609, <http://dx.doi.org/10.1038/s41467-022-35295-1>.
- [15] V.B. Vallevik, A. Babic, S.E. Marshall, S. Elvatun, H.M. Brøgger, S. Alagaratnam, B. Edwin, N.R. Veeraragavan, A.K. Befring, J.F. Nygård, Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare, *Int. J. Med. Informatics* 185 (2024) 105413, <http://dx.doi.org/10.1016/j.ijmedinf.2024.105413>.
- [16] B. Belgodere, P. Dognin, A. Ivankay, I. Melnyk, Y. Mroueh, A. Mojsilovic, J. Navratil, A. Nitsure, I. Padhi, M. Rigotti, J. Ross, Y. Schiff, R. Vedpathak, R.A. Young, Auditing and generating synthetic data with controllable trust trade-offs, *IEEE J. Emerg. Sel. Top. Circuits Syst.* (2024) <http://dx.doi.org/10.1109/JETCAS.2024.3477976>.
- [17] P. Diggle, P. Heagerty, K.-Y. Liang, S.L. Zeger, Missing values in longitudinal data, in: *Analysis of Longitudinal Data*, second ed., Oxford University Press, 2002, pp. 282–318, <http://dx.doi.org/10.1093/oso/9780198524847.001.0001>.
- [18] Z. Zhang, C. Yan, T.A. Lasko, J. Sun, B.A. Malin, SynTEG: a framework for temporal structured electronic health data simulation, *J. Am. Med. Inform. Assoc.* 28 (3) (2021) 596–604, <http://dx.doi.org/10.1093/jamia/ocaa262>.
- [19] J. Li, B.J. Cairns, J. Li, T. Zhu, Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications, *Npj Digit. Med.* 6 (1) (2023) 98, <http://dx.doi.org/10.1038/s41746-023-00834-7>.
- [20] M.S. Haleem, A. Ekuban, A. Antonini, S. Pagliara, L. Pecchia, C. Allocca, Deep-learning-driven techniques for real-time multimodal health and physical data synthesis, *Electron. (Switzerland)* 12 (9) (2023) <http://dx.doi.org/10.3390/electronics12091989>.
- [21] J.L. Achterberg, M.R. Haas, M.R. Spruit, On the evaluation of synthetic longitudinal electronic health records, *BMC Med. Res. Methodol.* 24 (1) (2024) <http://dx.doi.org/10.1186/s12874-024-02304-4>.
- [22] W. Wang, T.-W. Pai, Enhancing small tabular clinical trial dataset through hybrid data augmentation: Combining SMOTE and WCGAN-GP, *Data* 8 (9) (2023) 135, <http://dx.doi.org/10.3390/data8090135>.
- [23] B. Theodorou, C. Xiao, J. Sun, Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model, *Nat. Commun.* 14 (1) (2023) 5305, <http://dx.doi.org/10.1038/s41467-023-41093-0>.
- [24] G. Nikolentzos, M. Vazirgiannis, C. Xypolopoulos, M. Lingman, E.G. Brandt, Synthetic electronic health records generated with variational graph autoencoders, *Npj Digit. Med.* 6 (1) (2023) <http://dx.doi.org/10.1038/s41746-023-00822-x>.
- [25] I. Isasa, M. Hernandez, G. Epelde, F. Londoño, A. Beristain, X. Larrea, A. Alberdi, P. Bamidis, E. Konstantinidis, Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis, *BMC Med. Inform. Decis. Mak.* 24 (1) (2024) <http://dx.doi.org/10.1186/s12911-024-02427-0>.
- [26] L. Mosquera, K. El Emam, L. Ding, V. Sharma, X.H. Zhang, S.E. Kababji, C. Carvalho, B. Hamilton, D. Palfrey, L. Kong, B. Jiang, D.T. Eurich, A method for generating synthetic longitudinal health data, *BMC Med. Res. Methodol.* 23 (1) (2023) <http://dx.doi.org/10.1186/s12874-023-01869-w>.
- [27] H. Sun, H. Lin, R. Yan, Collaborative synthesis of patient records through multi-visit health state inference, 2023, <http://dx.doi.org/10.48550/arXiv.2312.14646>, arXiv preprint.
- [28] B. Theodorou, S. Jain, C. Xiao, J. Sun, ConSequence: Synthesizing logically constrained sequences for electronic health record generation, *Proc. AAAI Conf. Artif. Intell.* 38 (14) (2024) 15355–15363, <http://dx.doi.org/10.1609/aaai.v38i14.29460>.
- [29] C. Pang, X. Jiang, N.P. Pavinkurve, K.S. Kalluri, E.L. Minto, J. Patterson, L. Zhang, G. Hripscak, G. Gürsoy, N. Elhadad, K. Natarajan, CEHR-GPT: Generating electronic health records with chronological patient timelines, 2024, <http://dx.doi.org/10.48550/arXiv.2402.04400>, arXiv preprint.
- [30] P. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, Exploring longitudinal data, in: *Analysis of Longitudinal Data*, second ed., Oxford University Press, 2002, pp. 39–52, <http://dx.doi.org/10.1093/oso/9780198524847.001.0001>.
- [31] Z. Wang, J. Sun, PromptEHR: Conditional electronic healthcare records generation with prompt learning, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022, pp. 2873–2885, <http://dx.doi.org/10.18653/v1/2022.emnlp-main.185>.
- [32] A.S. Hashemi, K. Etmiani, A. Soliman, O. Hamed, J. Lundström, Time-series anonymization of tabular health data using generative adversarial network, IJCNN, in: 2023 International Joint Conference on Neural Networks, vol. 2023-June, IEEE, 2023, pp. 1–8, <http://dx.doi.org/10.1109/IJCNN54540.2023.10191367>.
- [33] N.I.-H. Kuo, M.N. Polizzotto, S. Finfer, F. Garcia, A. Sönnnerborg, M. Zazzi, M. Böhm, R. Kaiser, L. Jorm, S. Barbieri, The health gym: synthetic health-related datasets for the development of reinforcement learning algorithms, *Sci. Data* 9 (1) (2022) 693, <http://dx.doi.org/10.1038/s41597-022-01784-7>.
- [34] N.I.-H. Kuo, F. Garcia, A. Sönnnerborg, M. Böhm, R. Kaiser, M. Zazzi, M. Polizzotto, L. Jorm, S. Barbieri, Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV, *J. Biomed. Inform.* 144 (2023) 104436, <http://dx.doi.org/10.1016/j.jbi.2023.104436>.
- [35] M.P. Wand, M.C. Jones, Kernel Smoothing, 1995, <http://dx.doi.org/10.1007/978-1-4899-4493-1>.
- [36] P.J. Diggle, A.P. Verbyla, Nonparametric estimation of covariance structure in longitudinal data, *Biometrics* 54 (2) (1998) 401, <http://dx.doi.org/10.2307/3109751>.
- [37] E. Gringarten, C.V. Deutsch, Teacher's aide variogram interpretation and modeling, *Math. Geol.* 33 (4) (2001) 507–534, <http://dx.doi.org/10.1023/A:1011093014141>.
- [38] H.W. Kuhn, The Hungarian method for the assignment problem, *Nav. Res. Logist. Q.* 2 (1–2) (1955) <http://dx.doi.org/10.1002/nav.3800020109>.
- [39] N. Tomizawa, On some techniques useful for solution of transportation network problems, *Networks* 1 (2) (1971) <http://dx.doi.org/10.1002/net.3230010206>.

- [40] J. Edmonds, R.M. Karp, Theoretical improvements in algorithmic efficiency for network flow problems, *J. ACM* 19 (2) (1972) <http://dx.doi.org/10.1145/321694.321699>.
- [41] A. Johnson, T. Pollard, R. Mark, MIMIC-III clinical database (version 1.4), *PhysioNet* (2016) <http://dx.doi.org/10.13026/C2XW26>.
- [42] A.E. Johnson, T.J. Pollard, L. Shen, L.W.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, *Sci. Data* 3 (2016) <http://dx.doi.org/10.1038/sdata.2016.35>.
- [43] O. Gottesman, J. Futoma, Y. Liu, S. Parbhoo, L.A. Celi, E. Brunskill, F. Doshi-Velez, Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions, in: H.D. III, A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 2020, pp. 3658–3667, <http://dx.doi.org/10.5555/3524938.3525281>.
- [44] A. Aristodimou, A. Diavastos, C.S. Pattichis, A fast supervised density-based discretization algorithm for classification tasks in the medical domain, *Health Inform. J.* 28 (1) (2022) <http://dx.doi.org/10.1177/14604582211065397>.
- [45] Y. Wu, W. Wu, D. Yang, C. Xu, Z. Li, Neural response generation with dynamic vocabularies, in: *32nd AAAI Conference on Artificial Intelligence*, AAAI 2018, 2018, <http://dx.doi.org/10.1609/aaai.v32i1.11943>.
- [46] M.H. Huque, J.B. Carlin, J.A. Simpson, K.J. Lee, A comparison of multiple imputation methods for missing data in longitudinal studies, *BMC Med. Res. Methodol.* 18 (1) (2018) 168, <http://dx.doi.org/10.1186/s12874-018-0615-6>.
- [47] R.J. Little, A test of missing completely at random for multivariate data with missing values, *J. Amer. Statist. Assoc.* 83 (404) (1988) <http://dx.doi.org/10.1080/01621459.1988.10478722>.
- [48] C. Li, Little's test of missing completely at random, *Stata J.* 13 (4) (2013) <http://dx.doi.org/10.1177/1536867x1301300407>.
- [49] T. Adams, C. Birkenbihl, K. Otte, H.G. Ng, J.A. Rieling, A.F. Näher, U. Sax, F. Prasser, H. Fröhlich, On the fidelity versus privacy and utility trade-off of synthetic patient data, *IScience* 28 (2025) <http://dx.doi.org/10.1016/j.isci.2025.112382>.
- [50] M. Hernandez, P.A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, N. Aginako, Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees, *Front. Digit. Health* 7 (2025) <http://dx.doi.org/10.3389/fgth.2025.1576290>.
- [51] L. Pilgram, F.K. Dankar, J. Drechsler, M. Elliot, J. Domingo-Ferrer, P. Francis, M. Kantarcioglu, L. Kong, B. Malin, K. Muralidhar, P. Myles, F. Prasser, J.L. Raisaro, C. Yan, K.E. Emam, A consensus privacy metrics framework for synthetic data, *Patterns* 6 (2025) <http://dx.doi.org/10.1016/j.patter.2025.101320>.