Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Statistical Similarity: A Clinical Validation Framework for Synthetic Acute Myeloid Leukemia Cohorts Run head title: Clinical Validation of Synthetic AML Data

Adriano Zaghi
adriano.zaghi2@unibo.it
University of Bologna
Stefano Polizzi
University of Bologna
Enrico Giampieri
University of Bologna
Eric Sträng
Charité - Universitätsmedizin Berlin
Lars Bullinger
Charité - Universitätsmedizin Berlin
Gastone Castellani
University of Bologna

Article

Keywords:

Posted Date: September 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10435731/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose:

Access to high-quality clinical data is constrained by privacy regulations and institutional governance, limiting collaborative research and reproducible clinical AI. Although synthetic data have emerged as a promising solution, their adoption depends on rigorous validation of statistical fidelity, clinical utility, and privacy protection. Here, we propose a clinical validation framework and evaluate whether synthetic acute myeloid leukemia cohorts can support clinically meaningful analyses and improve survival prediction through training-set augmentation.

Material and Methods:

We developed a tailored Variational Autoencoder (VAE) to generate synthetic AML cohorts from the UK NCRI cohort, integrating genomic, cytogenetic, and longitudinal clinical data. Synthetic cohorts were evaluated for statistical fidelity, clinical utility, and privacy protection using a DOMIAS membership-inference attack. We then assessed whether they reproduced clinically meaningful analytical workflows—including molecular clustering, survival analysis, and multistate modeling—and whether synthetic cohorts improved survival prediction on independent real-world patients.

Results:

The synthetic cohorts faithfully reproduced the genomic, clinical, and longitudinal characteristics of the original AML cohort, including mutational profiles, molecular clustering, survival patterns, and multistate disease trajectories. Privacy assessment showed no evidence of patient-level leakage under a DOMIAS membership-inference attack. Synthetic training cohorts improved survival prediction on independent real-world patients while increasing the stability of predictive performance.

Conclusions:

Synthetic AML cohorts generated using a tailored VAE can support clinically meaningful analytical workflows while preserving patient privacy. Under specific circumstances, synthetic patients can improve survival prediction on independent data and the stability of predictive models. More broadly, our validation framework provides a practical strategy for establishing trust in synthetic clinical data and facilitating their adoption in clinical AI and translational research.

Context summary

Key Objective:

To demonstrate that synthetic multimodal AML datasets integrating NGS mutations, cytogenetic abnormalities, and longitudinal clinical outcomes can reproduce clinically meaningful analyses while preserving patient privacy and improving survival prediction on independent real-world patients.

Knowledge Generated:

Synthetic cohorts generated using a tailored variational autoencoder faithfully reproduced the genomic, clinical, and longitudinal characteristics of the original AML cohort, including mutational profiles, molecular clustering, survival patterns, and multistate disease trajectories, while showing no evidence of patient-level leakage under membership-inference attacks. When used for training-set augmentation, synthetic patients improved survival prediction on independent real-world patients and increased the stability of predictive performance.

Introduction

Advancements in medical research and clinical artificial intelligence (AI) rely heavily on access to high-quality patient data. However, acquiring such data is costly, harmonizing datasets across institutions is often difficult, and the sensitive nature of patient information imposes strict privacy regulations that limit data sharing and hinder collaborative research. Synthetic datasets offer a potential solution [1,2,3]. Synthetic datasets—artificial patient-level records generated by statistical or machine-learning models—have emerged as a promising approach to overcome these barriers.

These models learn the joint distribution of patient-level information from heterogeneous sources and generate new samples that resemble the original data without corresponding to any individual patient. As a result, synthetic data can be shared more freely across the scientific community while avoiding many of the regulatory and ethical constraints associated with real patient data.

To benefit from this technology, rigorous validation is essential. Most existing validation approaches emphasize statistical similarity between real and synthetic data, whereas considerably less attention has been devoted to validating their ability to support clinically meaningful analyses. Synthetic datasets must accurately reproduce the statistical and clinical properties of the original data, while also meeting acceptable privacy standards, since synthetic data are not intrinsically immune to re-identification risks. When properly validated, synthetic data can serve as surrogates, or even complements, to real data in downstream applications such as predictive modeling, clustering, and survival analysis. Recent work has also suggested that synthetic data may improve predictive performance by increasing the effective training set size and enhancing significant patterns, thus allowing the predictive model to better improve its performances.

A rigorous validation process is therefore required before synthetic data can be used in biomedical research. Validation should establish not only that synthetic data faithfully represent the statistical and clinical characteristics of the source population, but also that they provide adequate protection against patient re-identification. Once these requirements are satisfied, synthetic cohorts can support downstream analyses such as clustering, survival modeling, and disease-trajectory analysis. In addition, replacing real training cohorts with synthetic patients has recently emerged as a promising strategy for improving the robustness and predictive performance of clinical AI models.

Cancer provides an ideal setting for evaluating synthetic clinical data because robust AI models require large cohorts that capture the substantial biological and clinical heterogeneity of individual tumor entities. Acute myeloid leukemia (AML) represents a particularly demanding benchmark because prognosis depends on the integration of genomic alterations, cytogenetic abnormalities, and longitudinal disease trajectories, making faithful synthetic data generation especially challenging. In this study, we present a clinical validation framework based on replicating domain-specific analyses in synthetic data. We generated synthetic multimodal AML cohorts from the UK-NCRI AML[4,5,6,7] dataset using a tailored implementation of a Variational Autoencoder (VAE)[8,9], after observing that several open-source synthetic data generators were unable to reproduce basic statistical characteristics of the cohort. The resulting synthetic datasets are then evaluated using a clinically meaningful analytical workflow that includes clustering, survival analysis, and multistate model fitting. For survival analysis, we further evaluated whether synthetic cohorts improved prediction on an independent real-world validation cohort, with respect to the real data used for generating mentioned synthetic data. By comparing results across these complementary analyses, we assess the fidelity and limitations of the synthetic cohorts. We demonstrate that task-specific validation, combined with privacy assessment, provides a practical framework for establishing trust in synthetic clinical data. The synthetic cohorts were evaluated along three complementary dimensions: (i) privacy preservation, (ii) preservation of clinically meaningful disease structure, and (iii) utility for downstream predictive modeling.

Methods

In this section we describe the original data used to train the VAE, outline the implementation of the generative model, and detail the procedures used to assess privacy risk. We also report the statistical and clinical analyses, including clustering, survival modelling, and multistate model fitting, used to evaluate the statistical fidelity, clinical utility, and privacy of the synthetic cohorts.

Real AML cohort: UK-NCRI

The real dataset used to train our model originates from the UK-NCRI cohort, as compiled and annotated by Tazi et al. [4] to which we refer for full details on data acquisition and curation. The dataset used in this study is provided in tabular form and includes 2,017 patients. Each patient is represented by 70 recurrent cytogenetic abnormalities (frequency > 1%) and 84 recurrent gene mutations, each encoded as a binary variable indicating presence or absence.

In addition to baseline genomic information, each patient is described by six time-to-event variables representing transitions across possible disease states. The recorded time-to-event outcomes include overall survival, complete remission, relapse, death without remission, death in remission, and death in relapse. Together, these variables define longitudinal patient trajectories that can be represented within a multistate framework. Following Tazi et al.[4] We adopted a Markovian multistate model, assuming that transition intensities depend solely on a patient's current state rather than prior clinical history. Additional details, the full transition matrix of the multistate model, are provided in Supplementary Section 1.

Tailored VAE model

We implemented a Variational AutoEncoder (VAE)[8,9] to generate synthetic AML cohorts integrating genomic, cytogenetic, and longitudinal clinical information. Real-valued variables were reconstructed using a softmax-based output layer, whereas binary variables were reconstructed using a 0.5 threshold.

The model structure was tailored to accommodate the mixed genomic, cytogenetic, and longitudinal time-to-event variables of the dataset. The encoder mapped each patient record into a low-dimensional latent representation, parameterized by mean and variance vectors. The decoder then reconstructed patient-level features from samples drawn from the latent distribution.

A key methodological innovation of our approach is the use of long-format encoding for longitudinal survival data. Although long-format representations are routinely used for fitting multistate survival models, to our knowledge they have not previously been incorporated into the training of generative models for synthetic clinical data. Encoding patient trajectories in this form substantially reduced the generation of synthetic patients with biologically implausible state transitions, thereby improving the clinical consistency of the generated cohorts. This design choice was essential for preserving disease trajectories in the downstream multistate analyses.Additional technical details are provided in Supplementary Section 2.

Privacy risk assessment:

The primary privacy risk associated with synthetic data is the possibility that a generative model memorizes individual training records. We therefore evaluated privacy using Membership-Inference Attacks (MIAs), which aim to determine whether a specific individual was included in the model training set[10]. We implemented density estimators to compare the distributions of real and synthetic data, enabling the identification of potential overfitting—a known issue in complex neural network models that can increase privacy risk. To simulate a MIA, the generative model is trained on a subset of the real data (R_{train}) and used to generate two sets of synthetic samples (S and S_{test}). The attacker is assumed to have access..to all the synthetic data (S + Stest), as well as some background knowledge of the real-data distribution, obtained through publicly available datasets of the same type (R□□□□).

In our analysis, we focus on the mutational profiles of each patient. Dimensionality was reduced to two principal components using PCA before using the DomiasMIAKDE[11] class from the Synthcity package[12] that implements MIA using kernel density estimators.

Gaussian kernel density estimators (KDEs) are used to estimate the distributions of S and R□□□□, yielding density functions d(S) and d(R□□□□). The attacker attempts to classify the union of R□$_{r}$□$_{i}$□ and S□□□□ by examining a monotonic function of the density ratio d(S)/d(R□□□□). Regions where this ratio is greater than 1 (d(S)/d(R□□□□) > > 1) indicate potential over-representation of specific profiles in the synthetic data, suggesting that the generative model may have memorized training-set patients. Based on this ratio, the attacker constructs a classifier to infer whether a given sample belongs to R□$_{r}$□$_{i}$□.

Evaluating this classifier on (R□$_{r}$□$_{i}$□ + S□□□□) yields accuracy and AUC values, where lower accuracy corresponds to better privacy protection. This analysis quantifies whether the synthetic cohorts retain identifiable information about individual patients contained in the training data. Throughout the study, privacy assessment was considered jointly with statistical fidelity and clinical utility, reflecting the three complementary dimensions of trustworthy synthetic data.

Survival analysis

We evaluated whether synthetic cohorts preserved predictive information relevant for survival modelling and whether they could improve prediction through training-set augmentation. The primary endpoint of the survival analysis was generalization performance on previously unseen real patients. Survival models were trained on real and synthetic training cohorts and evaluated exclusively on independent held-out real patients that were never used during generative model training.[13] Survival models were fitted using penalized Cox proportional hazards regression with elastic-net regularization (lifelines, CoxPHFitter)[14]. The elastic-net mixing parameter was fixed at 0.5, whereas the regularization strength was optimized by cross-validation. To avoid information leakage, synthetic cohorts were generated independently within each training partition of the cross-validation procedure. Consequently, the independent held-out real patients used for model evaluation were never used either to train the VAE or to fit the survival model. This design ensured a strict separation between synthetic data generation, model development, and model evaluation.

Clustering with Dirichlet multistate model

Over the past decade, the Hierarchical Dirichlet Mixture Model (HDMM) [4, 15, 16, 17] has become a widely adopted Bayesian frameworkfor identifying molecular subtypes in onco-hematology. This Bayesian model automatically determines the optimal number of components (clusters) and provides a probabilistic characterization of each component. Each component is characterized by a vector of feature weights that defines its molecular signature. Mutations with the highest weights typically correspond to putative driver alterations, whereas individual patients are assigned deterministically to the component with the highest posterior probability.HDMM was selected because it represents the current reference approach for molecular stratification of AML and therefore provides a clinically meaningful benchmark for evaluating synthetic cohorts.Our implementation of HDMM Our implementation closely follows that of Tazi et al.[4] The model was fitted using 10 independent MCMC chains.For each chain, we discarded the first 5000 iterations to ensure that each chain has converged. The outcome of the 10 MCMCs was merged, collapsing components with a cosine similarity greater than 0.95. Mutations with fewer than five occurrences were excluded, as rare events are known to introduce convergence issues in the HDMM. Patients with no informative mutations were assigned to a “no useful mutation” cluster. Each resulting component is represented by a vector of probabilities (weights) over the genetic features. To obtain the clusters, each patient is assigned to the highest projection value component. Patients with identical posterior probabilities across components were assigned to an "Unclassified" cluster.Because HDMM determines the number of clusters directly from the data, independently fitted real and synthetic cohorts contained different numbers of components. Correspondence between components was therefore established through a dedicated alignment procedure described in Supplementary Section 4.

Results

Synthetic Data Generation and Model Selection

To generate synthetic data, we trained VAE models on the UK NCRI cohort until convergence of the reconstruction loss., (e.g.when the loss function reached a plateau). During training, we monitored synthetic-data quality using marginal distributions, mutation-per-patient patterns, and survival-time distributions. These metrics were also used to compare our VAE with two open-source models based on Generative Adversarial Networks (GAN models: CopulaGAN and CGAN), which failed to reproduce several key statistical and clinical characteristics of the real cohort. In contrast, the VAE reproduced all the evaluated data characteristics with close agreement with the real cohort, although some fluctuations around the optimum were observed. To mitigate this variability, we trained ten independent VAE models with identical architecture but different random seeds. For privacy and clustering analysis, we used either all ten synthetic cohorts or the cohort generated by the best-performing model, selected based on marginal-distribution similarity. For survival analysis, a different training strategy was used in order to isolate real testing patients. Additional details on convergence, benchmarking, and model selection are provided in Supplementary Section 3. These results indicate that the tailored VAE was substantially more capable than GAN-based alternatives at learning the complex joint distribution of the AML cohort, supporting its use as the backbone of the synthetic-data pipeline.

Privacy Validation

We evaluated MIA performance using the DOMIAS KDE-based approach. For each of the ten VAE models, synthetic data were split into *S* and *S*□□□□, while the held-out validation cohort served as *R*□□□□. KDE models were then trained on *S* and *R*□□□□ and used to classify the union of *S*□□□□ and *R*□$_{r}$□$_{i}$□. Across ten synthetic cohorts generated by the ten models that we trained, the mean AUC was 0.507 (SD = 0.009) and the mean accuracy was 0.504 (SD = 0.002).

Membership-inference performance was indistinguishable from random guessing, suggesting that the synthetic data do not reveal training-set membership and therefore meet acceptable privacy safeguards.

Survival Analysis

We evaluated survival prediction performance on different training cohorts, real and synthetic (generated using the VAE) ones. For each cohort, five Cox models were trained using five-fold cross-validation. When synthetic data were used to train the Cox models, they were generated by generative models trained exclusively on the four training folds of the real dataset. The Cox model trained on the resulting synthetic data was then evaluated on the held-out real data fold. The performance results are summarized in Fig. 1. Using training data composed by synthetic patients consistently improved survival prediction on independent held-out real patients. Since we are using synthetic data, cohort size is not constrained by the availability of real patients. Therefore, we evaluated the predictive potential of synthetic cohorts containing 10 times and 100 times more patients than the original real training cohort, and we observed a further increase in the C-index. Models trained on synthetic data systematically outperformed those trained on real data when the two cohorts had the same size. As the size of the synthetic cohort increased, we observed a further improvement in predictive performance (10× synthetic cohort). However, this improvement did not appear to be unlimited; instead, it plateaued when the cohort size was increased by another order of magnitude (100× synthetic cohort). These findings suggest that increasing synthetic cohort size yields diminishing returns beyond a certain point. Under these conditions, the average predictive performance was comparable to that observed with the 10× cohort, although we also observed a noticeable increase in the variability across models trained on different fold partitions.

Represents predictive accuracy (C-index) evaluated from Cox models trained on different train sets and tested on held out real data. We trained survival models on real data (excluding the subset of them used as test set), on fully synthetic data (in the same amount as the real train set) and on larger synthetic cohorts (10 and 100 times the real one). Cox models have been trained using a 5 fold cross validation that yields 5 different models for each training set, one for each training set partition. The same fold partition used for real data was used to determine the training sets for VAE models and the real held out test set for the Cox models trained on synthetic data obtained from VAEs. In this plot, we report the C-indexes obtained from testing each model on held out patients.

Dirichlet Clustering Consistency

We evaluated whether the synthetic cohorts preserved the molecular structure of AML by applying HDMM clustering. Real-data components were used as the reference, and both real and synthetic patients were assigned to these components (Fig. 2). Figure 2 shows that, across the ten synthetic cohorts generated by the ten models trained independently, cluster-membership variability remained within the synthetic standard deviation for all clusters except “No useful mutations,” which inherently displays higher heterogeneity.

Represents real (blue) and synthetic (orange) cohort distribution among components obtained fitting the HDMM on real data. For each cluster, we report the proportion it represents and, for the synthetic population, the standard deviation obtained by repeating the assignment with cohorts obtained from 10 different generative models trained with the same set up and different initial conditions.

Next, we fitted HDMM directly to synthetic data obtained from the best generative model and compared the resulting components (in this case 11 components, plus “unclassified” and “No useful mutations”) to those learned from the real cohort. Assigning real patients with real-derived and synthetic-derived components allowed us to establish clear correspondences between clusters (details in Supplementary Section 4 and Supplementary Fig. 6). The top five highest-weight mutations of each real component (Table 1) were predominantly preserved among the most weighted mutations of the corresponding synthetic components (Table 2). One exception was real Component 7, whose defining mutations were dispersed across several synthetic components. However, based on real patient assignment patterns, we identified synthetic Component 4 as its closest analogue. This is consistent with the observation that most patients in synthetic Component 4 originated from real Component 2, with which Component 7 shares several defining mutations. Patients left unclassified by real components are predominantly assigned to the synthetic component 6, while patients unclassified with synthetic components come mainly from real component 1 and 2 .Overall, the synthetic cohorts preserved the major molecular subtypes of AML, with discrepancies confined primarily to rare or highly heterogeneous components. Minor discrepancies occur for rare or heterogeneous components, but overall cluster fidelity is strong. Importantly, the preserved components corresponded to clinically established AML genomic subtypes, including NPM1/FLT3-mutated AML[18], TP53/complex-karyotype AML[19], CEBPA-bi mutated AML[20], RUNX1/SRSF2-related myeloid phenotypes[18], and core-binding factor leukemias characterized by inv(16) or t(8;21)[21]. These molecular subtypes are associated with distinct prognostic groups and therapeutic strategies, indicating that the synthetic cohorts preserve clinically actionable genomic structure.(add references). The defining genomic alterations of favorable-, intermediate-, and adverse-risk AML subtypes were consistently recovered in the synthetic cohorts (Tables 1 and 2).

Table 1
Real cohort most weighted mutations for each component obtained from HDMM fit on real data. Column colors match those of Table 2 in order to highlight components correspondence.

Real components	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8
Defining mutations	NPM1	SRSF2	complex	t(8;21)	inv(16)	CEBPA_bi	SF3B1	KMT2C
Defining mutations	DNMT3A	RUNX1	TP53	-Y	NRAS_p.Q61-62	GATA2	inv(3)	BAGE3
Defining mutations	ITD	TET2	− 5	KIT	KIT	TET2	− 7	GATA2
Defining mutations	TET2	ASXL1	− 7	− 9	KRAS	WT1	NRAS_p.G12-13	PTPN11
Defining mutations	IDH2_p.R140	DNMT3A	− 17	ASXL2	NRAS_ p.G12-13	KIT	PTPN11	KIT

Table 2
Most weighted mutations for each component obtained from HDMM fit on synthetic data. The top row of the table refers to Table 1 header, specifying the correspondence between real and synthetic components. Also table colors are meant to highlight the correspondence with Table 1. Synthetic component number 6 collects most of the patients which are considered "Unclassified" in real clustering. This can be seen in the Sankey plot in supplementary materials (Supplementary Fig. 6).

Associated real component	**Comp. 1**	**Comp. 2**	**Comp. 3**	**Comp. 2 e 7**	**Comp. 5**	**Unclassified**	**Comp. 2**	**Comp. 2**	**Comp. 4**	**Comp. 6**	**Comp. 8**
Synthetic component	**Comp. 1**	**Comp.2**	**Comp. 3**	**Comp. 4**	**Comp. 5**	**Comp. 6**	**Comp. 7**	**Comp. 8**	**Comp. 9**	**Comp. 10**	**Comp. 11**
Defining mutations	NPM1	SRSF2	complex	ASXL1	inv(16)	DNMT3A	DNMT3A	+ 8	t(8;21)	CEBPA	BAGE3
Defining mutations	ITD	RUNX1	TP53	RUNX1	+ 8	RUNX1	IDH2_p.R172	complex	-Y	GATA2	KMT2C
Defining mutations	DNMT3A	TET2	− 5	SRSF2	KIT	MLL	EZH2	+ 21	− 9	WT1	SF3B1
Defining mutations	PTPN11	ASXL1	− 7	SETBP1	FLT3_TDK	ITD	RUNX1	RUNX1	ASXL2	− 9	t(1;16)
Defining mutations	IDH2_p.R140	STAG2	− 17	TET2	NRAS_p.Q6162	U2AF1_p.S34	ASXL1	DNMT3A	KIT	NFE2	inv(3)

Multistate Model Comparison

To assess whether the synthetic cohorts preserved longitudinal disease trajectories, we compared multistate models fitted to the real cohort with those fitted to the synthetic cohort generated from the best-performing VAE. Cumulative baseline hazards for all transitions are shown in Fig. 3.

To test for systematic deviations, we pooled real and synthetic data and performed, for each transition, a likelihood-ratio test comparing a Cox model (Coxph [22, 23, 24]) with no covariates to a model including a covariate indicating data origin (real vs. synthetic). All likelihood-ratio tests were non-significant (all $P > > 0.05$).

These findings indicate that the synthetic cohorts faithfully preserve longitudinal disease dynamics, supporting their use for trajectory-based clinical analyses.. Preserving multistate dynamics is substantially more demanding than reproducing marginal variable distributions because it requires maintaining the temporal relationships among clinical events.

Represents cumulative baseline hazard of the five transitions with a confidence interval of 95%. Blue and orange curves represent results obtained from, respectively real and synthetic data. In the bottom right panel, we represent the multistate transition map.

Discussion

This study presents a clinical validation framework for evaluating synthetic patient cohorts and demonstrates its application to acute myeloid leukemia. Using this framework, we show that synthetic cohorts generated with a tailored VAE preserve clinically meaningful genomic structure, longitudinal disease trajectories, and predictive information while maintaining acceptable privacy protection.

The comparison with synthetic data obtained by GAN-based models proved that the VAE model was more suitable for reproducing real data patterns. We subsequently evaluated the utility of the data generated by VAE through clustering, survival analysis, and multistate model fitting. Across all analyses, VAE synthetic data reproduced key statistical and clinical properties of the real cohort and the domain inference attack that was simulated yields re identification performances close to random guessing, proving privacy safety of the synthetic cohort.

The most important practical finding of this study is that synthetic patients cohorts improve survival prediction on independent real patients. In survival modelling, synthetic data trained models yielded, on average, higher C-index values than models trained on real data. For two of these training strategies (Synthetic Data and Synthetic Data ×10), we also observed a reduced variance in the C-index distribution across different fold partitions. Although the model trained on the 100× synthetic cohort outperformed the real-data model on average, its performance was considerably more unstable, showing much greater variability across fold partitions. When comparing the different fold partition results, we also observed a pattern common to all survival models: models trained on the real training set of fold partition 3, or on synthetic data generated from the corresponding training set, consistently yielded lower predictive performance on the real test set, compared to models associated with the other partitions. An equally important observation is that synthetic cohorts faithfully reproduced not only informative biological patterns but also biases present in the original training data, illustrating that generative models preserve both informative biological patterns and cohort-specific limitations. This observation highlights the importance of validating synthetic cohorts before downstream use rather than assuming that synthetic generation inherently improves data quality. We further investigated how increasing synthetic cohort size influenced predictive performance of Cox models on real data. When using a synthetic cohort of the same size as the real one, we observed an increase in the average C-index together with a substantial reduction in the standard deviation of its distribution. This improvement was mainly driven by the increase in predictive performance on the test set of partition 1. Increasing the size of the synthetic cohort by a factor of 10 further improved the average C-index while maintaining a similar standard deviation. This suggests that, to fully exploit the potential of generative models, synthetic cohorts should be larger than the original real cohort. This is likely due to the generalization capability of the generative model, which can generate new patients that are not present in the original training set but remain consistent with the underlying data distribution, thereby improving predictions on the held-out real test sets. Increasing the

synthetic cohort size by a factor of 100 did not produce any meaningful improvement in the average C-index, but substantially increased the variability in predictive performance across fold partitions. This suggests that predictive performance reaches a plateau beyond a certain synthetic cohort size, whereas its variance across fold partitions continues to increase. A plausible explanation is that not all additional synthetic patients generated by the model remain equally consistent with the underlying data distribution, particularly when the training set contains biases, as observed for partition 3. In such cases, the generative model may amplify these biases. This effect appears to be independent of cohort size, as partition 3 consistently produced the poorest results. However, the sensitivity of the Cox models to this bias appears to depend on the size of the synthetic cohort. In particular, the Cox model trained on the x100 synthetic data generated from partition 3 and evaluated on the corresponding real test set appears to be more strongly affected by the partition-specific bias, likely because the larger number of synthetic patients carrying this bias cannot be sufficiently mitigated by the regularization terms ($L1 + \alpha \times L2$).

Clustering analyses further supported the fidelity of the synthetic cohort. HDMM components derived from real data were able to group synthetic patients in patterns consistent with those observed in the original cohort, indicating preservation of biological structure. Moreover, components inferred directly from synthetic data highlighted sets of high-weight mutations that closely matched known driver alterations in the real cohort. These findings further support that synthetic data may contribute meaningfully to exploratory analyses, including subtype identification and driver gene discovery.

From a clinical perspective, preservation of AML molecular structure is particularly relevant because modern AML risk stratification heavily depends on recurrent genomic alterations. The synthetic cohorts retained major mutation-defined entities associated with favorable, intermediate, and adverse prognosis, including

including NPM1/FLT3-mutated AML[18], TP53/complex-karyotype AML[19], CEBPA-bi mutated AML[20], RUNX1/SRSF2-related myeloid phenotypes[18], and core-binding factor leukemias characterized by inv(16) or t(8;21)[21]. These entities are clinically actionable because they guide treatment intensity, transplant decisions, and prognostic assessment. The ability of synthetic data to preserve these relationships suggests that such cohorts may support exploratory biomarker analyses and methodological development without requiring unrestricted access to sensitive patient-level datasets.

Multistate modelling demonstrated additional concordance between real and synthetic data. Cumulative baseline hazards for all transitions were broadly compatible, and likelihood-ratio testing showed no significant differences between models that included versus excluded a covariate indicating data origin. These results imply that the synthetic cohort can reproduce clinically meaningful disease trajectories within a multistate framework.

Our results suggest that trustworthy synthetic clinical cohorts should be evaluated across three complementary dimensions: statistical fidelity, clinical utility, and privacy protection. Together, these findings illustrate that synthetic data can support complex analytical workflows, although with some limitations. Here we summarize the main strengths and limitations emerging from our analyses. First, the synthetic cohorts inherit the structure and potential biases of the dataset used for training the generative model. The effect of biases in synthetic data on downstream analyses can differ from their effect in real data when the size of the synthetic cohort is increased, as observed in the survival analysis. Second, synthetic data cannot capture biological patterns that are absent from the source population, but they can generalize patterns present in the training data, potentially improving downstream analyses, as illustrated by the survival analysis. Third, although the synthetic data reproduced the major genomic and survival characteristics, discrepancies remained for rare mutations and heterogeneous components, which are intrinsically difficult to model due to their limited representation in the training data. Finally, while privacy attacks yield performance close to random guessing, privacy preservation cannot be considered absolute and should always be evaluated alongside clinical fidelity. It is important to consider that our analyses were performed within a single AML cohort, and additional validation across independent datasets and clinical settings will be required to establish generalizability. Overall, our findings support a pragmatic view of synthetic data in translational oncology: synthetic cohorts should not be considered perfect substitutes for real clinical datasets, but rather complementary resources capable of reducing barriers related to data scarcity and privacy. They can be used for methodological development, to perform preliminary analyses, or to improve the robustness and precision of analyses based on real data.

In light of those considerations, we propose that synthetic data should be accompanied by a standardized validation report comparing real and synthetic analyses across key domains, enabling researchers to understand potential biases and decide whether the synthetic dataset is appropriate for their use case. In a realistic deployment scenario, such validation would ideally be performed by a centralized facility with access to the protected real data, ensuring transparency and safe downstream use by end-users. More generally, we envision that synthetic clinical cohorts should be distributed together with standardized validation reports documenting statistical fidelity, clinical utility, and privacy protection. Such reports would enable researchers to assess the suitability of synthetic cohorts for specific downstream applications and facilitate trustworthy data sharing across institutions.

Declarations

Founding declaration

This study did not receive any specific funding. However, the PhD fellowship of Adriano Zaghi was funded by the EU Horizon Europe Research and Innovation program under grant agreement No 101095530 “SYNTHEMA” project. The funding source had no role in the study design, data collection, analysis, interpretation of the results, writing of the manuscript, or decision to submit the manuscript for publication.

Author Contribution

A.Z. conceived and designed the study, implemented the models, performed the analyses, drafted the manuscript, and prepared figures and tables. S.P. contributed in designing the study, performed the analyses related to Dirichlet model-based clustering, and wrote the corresponding sections of the

manuscript. E.G. and contributed in designing the study together with the other authors. E.S. supervised the implementation of the VAE model and contributed in designing the study. L.B. and G.C. supervised and supported the whole project and contributed in designing the study. All authors reviewed and approved the final manuscript.

Acknowledgement

This work was supported by the AIRC Foundation (Fondazione AIRC per la Ricerca sul Cancro ETS; Milan, Italy; project No 26216 to GC) and by the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 116026 H2020 EU "HARMONY" project and under grant agreement No 945406 “HARMONY PLUS” project (to GC, CS, LB, ES and EG). This Joint Undertaking receives support from the European Union’s Horizon 2020 Research and Innovation program and EFPIA. This work was also supported by the EU H2020 “GenoMed4All” project (grant agreement No 101017549) (to GC, CS and EG) and by the EU Horizon Europe Research and Innovation program under grant agreement No 101095530 “SYNTHEMA” project (to GC, CS, LB, ES and EG) and by the EU IHI project SYNTHIA GA n. 101172872 (to GC, CS, LB, ES and EG). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union and the European Union can't be held responsible for them.

Data Availability

The primary real-world data utilized in this study are derived from the UK National Cancer Research Institute (NCRI) AML cohort, which was previously compiled and made available in the publication by Tazi et al.4. Access to these patient-level data is subject to the original data-sharing agreements and the governance framework of the UK-NCRI.To prioritize the sharing of reproducible methodologies over static data, we do not provide the individual synthetic datasets generated for these analyses. Instead, we have released the complete source code implementing the tailored Variational Autoencoder (VAE) and the associated validation framework. Detailed descriptions of our generative strategies—including the long-format encoding developed to preserve multistate disease trajectories—are provided within this manuscript and its supplementary materials to ensure full results reproducibility. The implementation code and the real data are available at: https://github.com/AdrianoZaghi/VAE_for_UKNCRI.git .

References

1. Aldren Gonzales, Guruprabha Guruswamy, Scott R. Smith. Synthetic data in health care: A narrative review PLOS Digital Health 2(1): e0000082
2. Giuffrè, M., Shung, D.L. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3
3. Boyeau, Pierre, et al. "Autoeval done right: Using synthetic data for model evaluation." *arXiv:2403.07008*(2024).
4. Yanis Tazi, Juan E Arango-Ossa.. Unified classification and risk-stratification in Acute Myeloid Leukemia, PMID: 35941135 PMCID: PMC9360033, DOI: 10.1038/s41467-022-32103-8
5. Burnett AK, Russell NH, Hills RK, et al. A randomized comparison of daunorubicin 90 mg/m2 vs 60 mg/m2 in AML induction: results from the UK NCRI AML17 trial in 1206 patients. *Blood*. 2015;125(25):3878-3885. doi:10.1182/blood-2015-01-623447
6. Burnett AK, Hills RK, Hunter AE, Milligan D et all. The addition of gemtuzumab ozogamicin to low-dose Ara-C improves remission rate but does not significantly prolong survival in older patients with acute myeloid leukaemia: results from the LRF AML14 and NCRI AML16 pick-a-winner comparison. Leukemia. 2013 Jan;27(1):75-81. doi:10.1038/leu.2012.229. PMID:22964882.
7. Burd, A. et al. Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the Beat AML Master Trial. Nat. Med. https://doi.org/10.1038/s41591-020-1089-8 (2020).
8. Kingma, Diederik P., and Max Welling. "An introduction to variational autoencoders." *Foundations and Trends® in Machine Learning* 12.4 (2019).
9. Hajra M. et all., Synthetic data generation: State of the art in health care domain, Computer Science Review, Volume 48, 2023, 100546, ISSN 1574-0137,
10. Membership inference attacks against machine learning models. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. In *2017 IEEE symposium on security and privacy (SP)* (2017, May).
11. Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M. (2023). Membership inference attacks against synthetic data through overfitting detection. *arXiv preprint arXiv:2302.12580*.
12. Qian, Z., Cebere, B.-C., & van der Schaar, M. (2023). Synthcity: Facilitating innovative use cases of synthetic data in different data modalities. arXiv preprint arXiv:2301.07573.
13. Akiya, I., Ishihara, T., & Yamamoto, K. (2024). Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: simulation study. *JMIR medical informatics*.
14. Davidson-Pilon, (2019). lifelines: survival analysis in Python. Journal of Open Source Software, 4(40), 1317, https://doi.org/10.21105/joss.01317.
15. Papaemmanuil E., Gerstung M., Bullinger L. et all. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England Journal of medicine*.
16. Dall'Olio D., Sträng E., Turki A. et all. Covering Hierarchical Dirichlet Mixture Models on binary data to enhance genomic stratifications in onco-hematology. *PLoS computational biology*.
17. Bersanelli M., Travaglino E., Meggendorfer M. et all. Classification and Personalized Prognostic Assessment on the Basis of Clinical and Genomic Features in Myelodysplastic Syndromes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. https://doi.org/10.1200/JCO.20.01659

18. E. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med*. 2016;374(23):2209-2221. doi:10.1056/NEJMoa1516192
19. Shahzad, M., Amin, M.K., Daver, N.G. et al. What have we learned about TP53-mutated acute myeloid leukemia?. Blood Cancer J. 14, 202 (2024). https://doi.org/10.1038/s41408-024-01186-5
20. Su L, Shi YY, Liu ZY, Gao SJ. Acute Myeloid Leukemia With CEBPA Mutations: Current Progress and Future Directions. Front Oncol. 2022;12:806137. Published 2022 Feb 1. doi:10.3389/fonc.2022.806137
21. Borthakur, G., Kantarjian, H. Core binding factor acute myelogenous leukemia-2021 treatment algorithm. *Blood Cancer J.* **11**, 114 (2021). https://doi.org/10.1038/s41408-021-00503-6
22. Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes, a large sample study. *Annals of Statistics* 10, 1100-1120.
23. Terry M. Therneau, Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York. ISBN 0-387-98784-3.
24. Therneau T (2024). *A Package for Survival Analysis in R*. R package version 3.8-3, https://CRAN.R-project.org/package=survival.

Figures

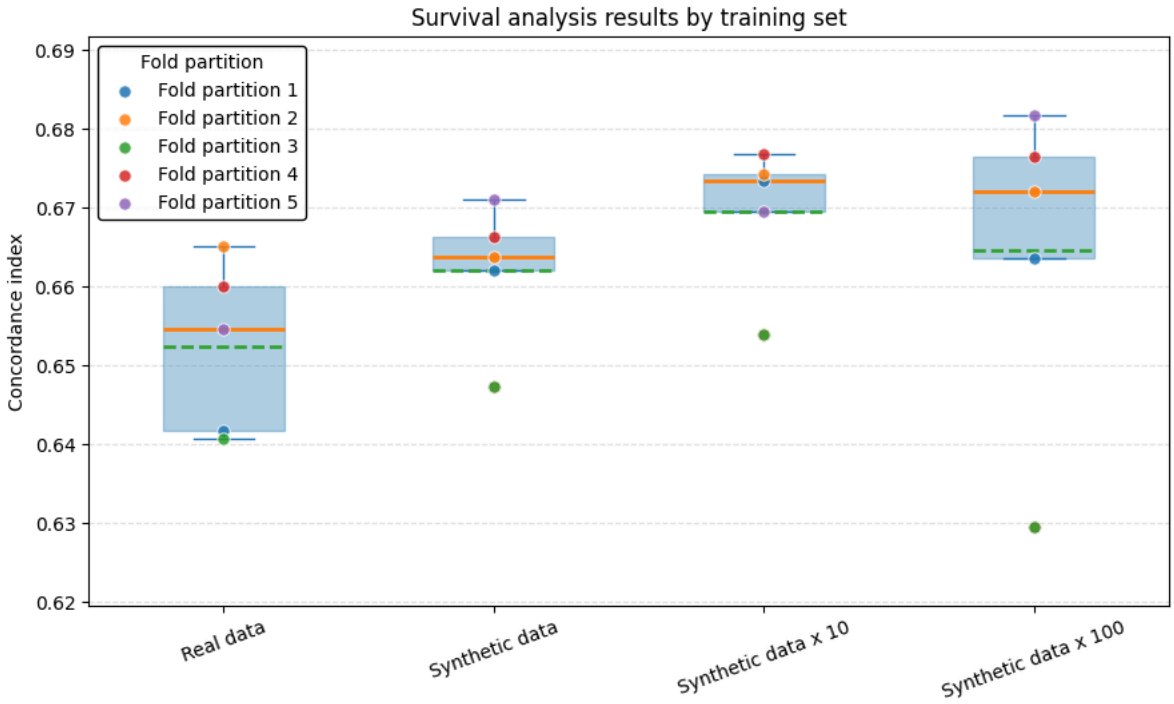


Figure 1

Represents predictive accuracy (C-index) evaluated from Cox models trained on different train sets and tested on held out real data. We trained survival models on real data (excluding the subset of them used as test set), on fully synthetic data (in the same amount as the real train set) and on larger synthetic cohorts (10 and 100 times the real one). Cox models have been trained using a 5 fold cross validation that yields 5 different models for each training set, one for each training set partition. The same fold partition used for real data was used to determine the training sets for VAE models and the real held out test set for the Cox models trained on synthetic data obtained from VAEs. In this plot, we report the C-indexes obtained from testing each model on held out patients.

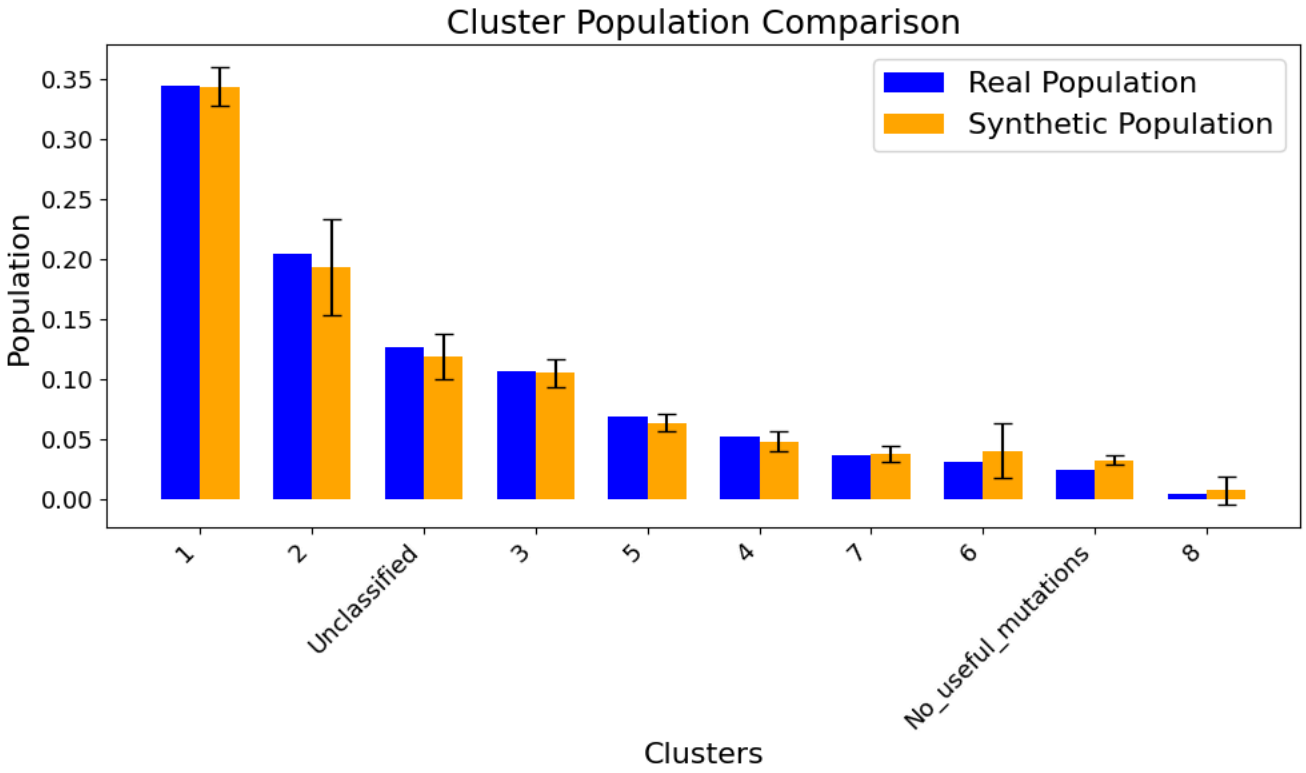


Figure 2

Represents real (blue) and synthetic (orange) cohort distribution among components obtained fitting the HDMM on real data. For each cluster, we report the proportion it represents and, for the synthetic population, the standard deviation obtained by repeating the assignment with cohorts obtained from 10 different generative models trained with the same set up and different initial conditions.

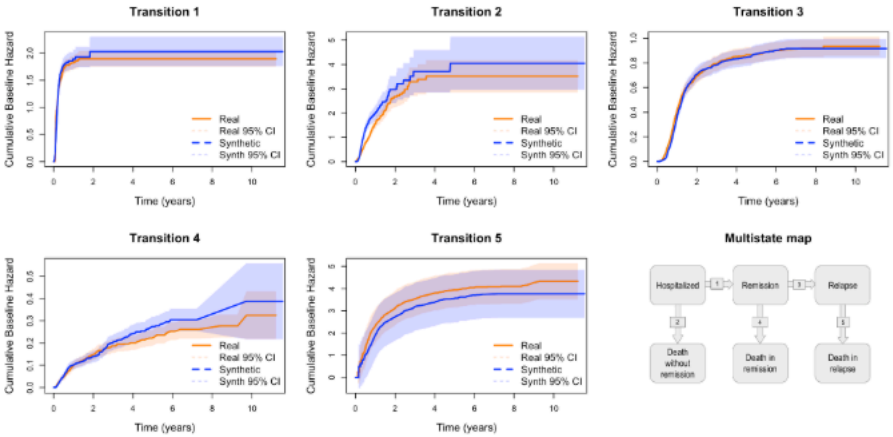


Figure 3

Represents cumulative baseline hazard of the five transitions with a confidence interval of 95%. Blue and orange curves represent results obtained from, respectively real and synthetic data. In the bottom right panel, we represent the multistate transition map.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx