Received: 12 February 2024 | Accepted: 13 February 2024

DOI: 10.1002/jha2.873

eJHaem

British Society for Haematology

RESEARCH ARTICLE

Unlocking the potential of synthetic patients for accelerating clinical trials: Results of the first GIMEMA experience on acute myeloid leukemia patients

Alfonso Piciocchi[1] | Marta Cipriani[1,2] | Monica Messina[1] | Giovanni Marconi[3] | Valentina Arena[1] | Stefano Soddu[1] | Enrico Crea[1] | Maria Valeria Feraco[4] | Marco Ferrante[4] | Edoardo La Sala[1] | Paola Fazi[1] | Francesco Buccisano[5] | Maria Teresa Voso[5] | Giovanni Martinelli[3] | Adriano Venditti[5] | Marco Vignetti[1]

[1]Data Center, GIMEMA Foundation, Rome, Italy

[2]Department of Statistical Sciences, University of Rome La Sapienza, Rome, Italy

[3]Hematology Unit, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", Meldola, Italy

[4]Department Health Care and Life Sciences, Studio Legale FLC, Rome, Italy

[5]Department of Biomedicine and Prevention, Tor Vergata University, Rome, Italy

Correspondence
Marta Cipriani, GIMEMA Foundation, Rome, Italy.
Email: m.cipriani@gimema.it

Abstract

Artificial Intelligence has the potential to reshape the landscape of clinical trials through innovative applications, with a notable advancement being the emergence of synthetic patient generation. This process involves simulating cohorts of virtual patients that can either replace or supplement real individuals within trial settings. By leveraging synthetic patients, it becomes possible to eliminate the need for obtaining patient consent and creating control groups that mimic patients in active treatment arms. This method not only streamlines trial processes, reducing time and costs but also fortifies the protection of sensitive participant data. Furthermore, integrating synthetic patients amplifies trial efficiency by expanding the sample size. These straightforward and cost-effective methods also enable the development of personalized subject-specific models, enabling predictions of patient responses to interventions. Synthetic data holds great promise for generating real-world evidence in clinical trials while upholding rigorous confidentiality standards throughout the process. Therefore, this study aims to demonstrate the applicability and performance of these methods in the context of onco-hematological research, breaking through the theoretical and practical barriers associated with the implementation of artificial intelligence in medical trials.

KEYWORDS

AML, machine learning, synthetic data, virtual patients

Alfonso Piciocchi and Marta Cipriani contributed equally to this work

1 | INTRODUCTION

Artificial Intelligence (AI) is revolutionizing numerous medical fields, ranging from image analysis and multi-omics data integration to drug discovery and precision medicine, with some notable applications already explored [1, 2]. Among these, clinical trial design emerges as a promising yet relatively unexplored frontier. While synthetic data has been originally designed to furnish publicly accessible alternatives to datasets, it holds substantial potential in clinical trial applications, providing a virtual patient group that closely emulates real-world datasets while safeguarding individual patient privacy. However, it is crucial to recognize that attaining these benefits demands a meticulous and resource-intensive process rather than inheriting them by default. This involves substituting observed values with sampled values from appropriate probability distributions, thereby preserving the statistical characteristics of the original data. By adopting this approach, the need for patient consent is alleviated, and limitations associated with common anonymization practices are overcome. Notably, ensuring the efficacy of anonymization has become increasingly challenging in the face of ongoing technological advancements [3].

In this context, our project was conceived to explore the feasibility and utility of synthetic data as a viable substitute for actual clinical trial data. One immediate application of synthetic datasets in clinical trials is the generation of control patient groups that faithfully mirror the characteristics of the original dataset. Generating synthetic patients for clinical trials can be accomplished through different methodologies, including Generative Adversarial Networks (GANs) [4], decision tree methods [5], and parametric methods. Decision tree methods, functioning as machine learning algorithms, sequentially classify or predict outcomes by partitioning data into subsets based on feature values. Parametric methods involve fitting a specific statistical distribution to observed data and then generating synthetic patients based on the estimated parameters of that distribution. In this work, we focus on decision tree and parametric methods, while an in-depth review and comparison with the GANs approach can be found in Little et al. [6].

Leveraging synthetic data enables the execution of virtual randomized trials, comparing the responses of patients receiving active therapy with those in a synthetic control group. However, the success of this approach requires empirical validation. Such an approach holds the promise of allowing investigators to strategically allocate resources by enrolling more patients in the active therapy arms, thereby optimizing trial efficiency and resource utilization.

2 | METHODS

2.1 | Study population

In the present study, we generated an in-silico [7] cohort using a sizable dataset of patients enrolled in the GIMEMA AML1310 study (NCT01452646), currently closed [8]. AML1310 was a trial for newly diagnosed AML patients which included a "3+7″-like induction and a risk-based MRD-directed post-remission transplant allocation. The study enrolled 500 patients across 55 Italian hematology institutions. Patients had a median age of 49 years, all of whom underwent an extensive biological characterization, including morphological, cytogenetic, molecular genetics, and multiparametric flow cytometry analyses. In particular, the individual-level data of the subset of 445 patients with ELN2017 risk classification available [9] was used to generate the synthetic cohort as described below.

2.2 | Statistical methods

2.2.1 | Model description

The model we used was designed to capture the pattern and statistical properties of the original data. Ideally, if the models truly represent the process that generated the original observed data, an analysis based on the synthesized data should lead to the same statistical inferences as an analysis based on the actual data. More in detail, in most implementations of synthetic data generation, the generative model uses a joint distribution made up of conditional distributions to produce synthetic data. Each column of the synthetic dataset is selected, with the distribution of that variable being estimated conditional on observed variables and all previous columns synthesized. This process is repeated for each subsequent column. The 'synthpop' R package [10] was utilized for this purpose, and Elliot [11] stated his agreement with the low disclosure risk associated with synthetic data produced using this package. The 'synthpop' package incorporates precautionary features, including a function that allows for top and bottom coding and the addition of labels to synthetic datasets to indicate their synthetic nature. Furthermore, this function excludes any unique cases with variable sequences identical to those of unique individuals in the real dataset from the synthetic dataset. In our specific case, it is worth noting that utilizing this function obviated the need to remove any units from our dataset. Thus, synthetic data derivatives are quantitatively the same as patient-derived datasets, but the former cannot be traced back to the individuals from whom they were derived.

2.2.2 | Choice of model parameters

The choice of synthesizing models typically involves a decision between parametric and non-parametric methods, with the latter being based on classification and regression trees (CART), which can accommodate any data type. In our study, we employed the parametric method for all variables except time-to-event variables, where the CART method by Breiman et al. [5] was used. Namely, the algorithm assigns default parametric methods to variables to be synthesized based on their types: for binary data type the logistic regression, for a factor with > 2 levels the polytomous logistic regression, for an ordered factor with > 2 levels the ordered polytomous logistic regression. For continuous variables, we employed a linear regression model that preserves the marginal distribution. Regression is conducted on an inverse normal-rank-based transformation of the covariate of interest to approximate continuous covariates with non-normal distributions. Additionally, given the algorithm structure, it is crucial to determine the order in which variables should be synthesized.

The use of the CART method for time-to-event variables and the parametric method for all other variables was motivated by the complex relationships often present in time-to-event data, often not easily captured by parametric models. The CART method, being a nonparametric technique, provides greater flexibility in modeling such intricate relationships, making it a suitable choice for these specific variables. Conversely, for the other variables in our study, the parametric method was chosen due to its simplicity and interpretability. Parametric models are often more straightforward to implement and understand, making them a practical choice when the relationships in the data are reasonably well-understood and conform to the assumptions of the chosen parametric model.

The variables synthesized included age, sex, height, weight, WHO performance status, neutrophil count, lymphocytes, hemoglobin, white blood cells (WBC), FLT3-ITD, NPM1, CBF, ELN2017 risk category, complete response (CR) rate, MRD-negativity, transplant rate, overall survival (OS), and disease-free survival (DFS).

2.2.3 | Data consistency and missing data

Before generating synthetic data, rules were implemented to ensure data consistency. Specifically, we identified restricted values, which are situations where certain values are explicitly determined by other variables. During the data synthesis process, restricted values are assigned first, followed by the generation of records with unrestricted values. According to the AML1310 study protocol, the rules applied include scenarios where patients not in complete remission do not provide measurements for MRD or DFS, and they have not undergone transplantation.

Missing data were handled differently for categorical and continuous variables. When dealing with missing data in categorical variables, the employed algorithm treats the missing values as additional categories, and replicating them is a straightforward process. For continuous variables that have missing data, a two-step modeling approach is employed. In the first step, an auxiliary binary variable is created to indicate whether a value is missing or not. In the second step, a synthesizing model is fitted to the non-missing values in the original variable. This model is then used to generate synthetic values for the non-missing category records in the auxiliary variable. Instead of using the original variable, the auxiliary variable along with a variable containing non-missing values and zeros is used for the remaining records when predicting other variables.

2.2.4 | Comparison between original and synthetic cohort

We used rigorous statistical methods to examine the faithfulness of the synthetic cohort to the original one.

To identify disparities in the distributions of the original and synthetic data and subsequently refine our synthesis methods for enhanced utility, we adopted an approach to general utility measures that involves combining the original and synthetic records. This method measures how well the data values predict the source of the records, distinguishing between real and synthetic, utilizing the propensity score—the predicted probability that a record originates from the synthetic data. Following the recommendation of Snoke et al. [12], the most commonly suggested utility measure in this context is the propensity score mean squared error (pMSE) or its standardized ratio (S_pMSE).

Drawing upon the works of Raab et al. [13], we propose a practical threshold for assessing utility: if all standardized pMSE ratios fall below 10, and preferably below 3, further adjustments may be deemed unnecessary, indicating the satisfactory utility of the synthetic data.

Alternatively, an additional approach to utility measures involves grouping the original and synthetic data by constructing tables based on their values and computing measures of difference between these tables. In our evaluation of adherence across both continuous and categorical variables, we employed the Wilcoxon rank sum test and Pearson's Chi-squared test. The survival outcomes of the virtual and actual cohorts were evaluated using the Log-rank test.

Recognizing that a high degree of statistical similarity may not necessarily imply true similarity, especially when considering the intricate interplay between variables, and appreciating the critical importance of these interrelationships—particularly in contexts like randomized controlled trials (RCTs), where subgroup analyses are essential—we conducted a thorough stratified survival analysis. This enabled us to delve into the influence of factors on distinct patient subgroups.

3 | RESULTS

3.1 | Virtual cohort characteristics

Implementing the described synthetic data generation approach yielded a virtual cohort comprising 890 patients, effectively doubling the size of the original sample. The covariate distributions found in the original data were replicated with a high degree of accuracy in the synthetic data, ensuring that survival predictions being made conditional on unique covariate patterns are appropriately reflected. Table 1 summarizes the features of the synthetic cohort and compares them to the original population. Notably, the clinic-biological characteristics of the two cohorts did not differ significantly, as all adjusted p-values exceeded 0.99.

To underscore these findings, Table S1 presents the obtained pMSE and standardized pMSE, along with their respective degrees of freedom. Remarkably, for every synthesized variable, the associated utility measure falls below the suggested threshold of 3 for the standardized pMSE.

TABLE 1 Comparison between the original and synthetic AML1310 cohorts in terms of demographic, clinic-biologic characteristics, and response.

Characteristic	Original AML1310 N = 445	Synthetic AML1310 N = 890	p-Value[a]	q-Value[b]
Age, median (range)	49 (18, 61)	49 (18, 61)	>0.99	>0.99
Sex, n (%)			0.97	>0.99
Male	232 (52%)	465 (52%)		
Female	213 (48%)	425 (48%)		
Height, median (range)	170 (80, 196)	169 (80, 196)	0.64	>0.99
Weight, median (range)	72 (43, 192)	71 (43, 192)	0.97	>0.99
WHO Performance Status, *n* (%)			0.51	>0.99
0	259 (60%)	535 (62%)		
1	120 (28%)	225 (26%)		
2	52 (12%)	88 (10%)		
3	2 (0.5%)	9 (1.1%)		
Neutrophils, median (range)	1.4 (0, 30)	1.3 (0, 30)	0.87	>0.99
Lymphocytes, median (range)	2.8 (0, 20)	3.0 (0, 20)	0.61	>0.99
Haemoglobin, median (range)	8.90 (3.30, 15.20)	8.90 (3.30, 15.20)	0.64	>0.99
WBC x 10^9, median (range)	14 (0, 341)	13 (0, 341)	0.69	>0.99
RUNX1, *n* (%)			0.69	>0.99
Negative	417 (94%)	835 (94%)		
Positive	27 (6.1%)	49 (5.5%)		
CBF, *n* (%)			0.25	>0.99
Negative	405 (92%)	791 (90%)		
Positive	36 (8.2%)	89 (10%)		
FLT3-ITD, *n* (%)			0.73	>0.99
Negative	334 (76%)	671 (76%)		
Positive	108 (24%)	207 (24%)		
NPM1, *n* (%)			0.62	>0.99
Negative	274 (62%)	558 (63%)		
Positive	170 (38%)	326 (37%)		
ELN2017 Risk category, *n* (%)			0.58	>0.99
Adverse	80 (18%)	174 (20%)		
Favorable	186 (42%)	383 (43%)		
Intermediate	179 (40%)	333 (37%)		
CR, *n* (%)	322 (73%)	645 (73%)	0.96	>0.99
MRD, *n* (%)			0.58	>0.99
Neg	125 (52%)	234 (49%)		
Pos	117 (48%)	239 (51%)		
Transplant received, *n* (%)	217 (49%)	432 (49%)	0.94	>0.99

[a] Wilcoxon rank sum test; Pearson's Chi-squared test.

[b] False discovery rate correction for multiple testing using the Bonferroni method.

Moreover, the *synthpop* package provides visualization methods for these results, including histograms comparing the original and synthetic distributions side by side (Figures S1–S17). This feature offers immediate feedback, enabling the data synthesizer to enhance the quality of the synthetic data.

3.2 | Virtual cohort outcomes

In terms of response evaluations, the synthetic cohort exhibited an aligned CR rate of 73%, perfectly consistent (q-value > 0.99) with the original cohort's CR rate. Furthermore, an exact concordance

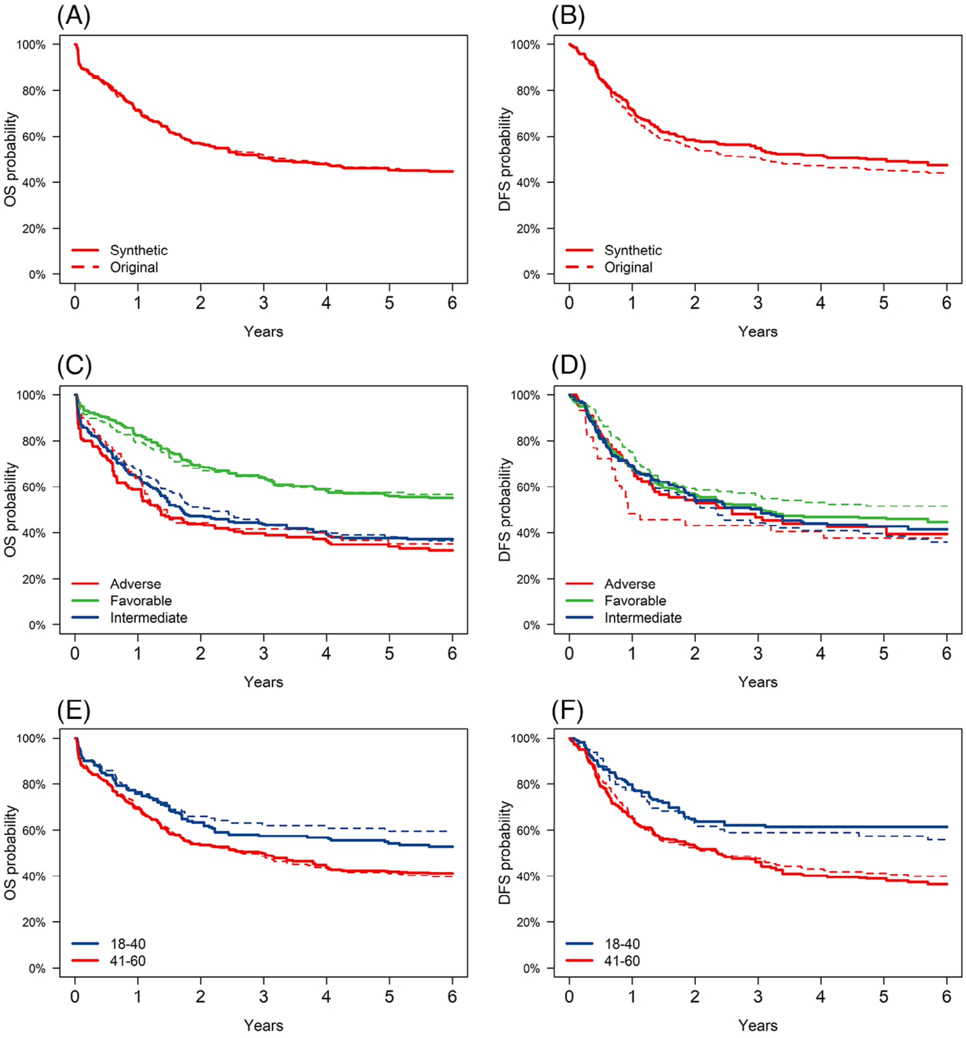

FIGURE 1 Survival outcomes of the original and synthetic AML1310 cohorts. Dashed lines refer to the original cohort and continuous lines to the synthetic patients.

(q-value > 0.99) was observed in MRD-negativity rates, with both cohorts recording a rate of 52%.

Furthermore, the effectiveness of these methods in replicating survival patterns found in the original data is demonstrated by comparing various survival estimates for the original and synthetic data. The marginal OS and DFS curves (Figure 1A,B) demonstrated striking similarity, with Log-rank test *p*-values of 0.78 and 0.79, respectively. Indeed, at 2 years OS was 57% (95% CI: 52.5%–61.9%) in the original cohort and 55.9% (95% CI: 52.7%–59.3%) in the synthetic AML1310 cohort. Similarly, two-year DFS was 55.1% (95% CI: 49.8%–60.9%) in the original and 55.9% (95% CI: 52.1%–60%) in the synthetic cohort. Remarkably, the synthetic curves for both OS and DFS managed to precisely reproduce the censoring distribution of the original population. This result allows us to avoid the imposition of assumption on the censoring independence from the time-to-event distribution.

Stratified survival estimates by population subgroups, presented in Figure 1 (plots C–F), further underscored the agreement in survival estimates over time. OS curves were stratified by risk categories (Figure 1C) and age class (Figure 1E), and equivalently DFS curves were stratified by risk categories (Figure 1D) and age class (Figure 1F).

4 | DISCUSSION

In this work, we demonstrated the feasibility of generating a virtual cohort of patients from real patients' data in the setting of AML. This study represents a concrete example of the implementation of AI in clinical trial design. Some experiences in the same or other settings have been recently published. Though employing different methods to generate synthetic patients, they witness the increasing interest and potential of synthetic data [14, 15].

In the present work, by employing innovative computational modeling techniques, we were able to develop an in-silico AML population whose main features are very similar to the real population. Mirroring an AML population treated with a conventional chemotherapeutic approach, the synthetic AML1310 cohort is suitable to represent the control group when testing novel innovative treatments, most likely in an in-silico randomized trial, performed in the same framework (i.e., the Italian Hematology Centres to ensure the usability of the synthetic population).

Indeed, while randomized controlled trials remain the gold standard for evaluating the safety and efficacy of new treatments, there is a mounting recognition of the need for alternative approaches to expedite the trial process.

Besides the abovementioned potential, we can also count some benefits for patients. Using a synthetic cohort generated from a conventionally treated population as the control group, the patients enrolled in the virtual randomized trial would receive only the experimental treatment without being exposed to the “less active” therapy, thus limiting treatment failures and toxicity.

Shifting to an in-silico trial would also be advantageous for all the clinical trial stakeholders: indeed, by reducing the need to enroll additional physical patients, enrolment and the attainment of final results would be faster, and investigation-related expenses would be optimized.

Another advantage is the privacy safeguard: indeed, completely synthesized data does not include identifiable real units, hence, the probability of disclosing a person’s identity is considered to be unlikely. Despite Rajotte et al. [16] consideration of the trade-off between ‘high-quality’ synthetic data creation and privacy issues, the synthpop R method—used in the present study—is designed to enhance data security and further minimize the potential risk of disclosure [10, 11].

In addition, the “burden” of collecting data subject’s consent as well as the shortcomings of common anonymization techniques are reduced.

Furthermore, by generating synthetic patients, one can address the limitations of small sample sizes or imbalances in covariates. This is particularly beneficial in propensity score matching, as having a more balanced set of covariates enhances the accuracy and reliability of treatment effect estimates. Especially in cases where certain events or conditions are rare in real-world data, synthetic patient generation can help create instances of these rare events, making it easier to match treated and control groups on such variables. Moreover, synthetic patient generation is a valuable tool for upsampling minority classes in imbalanced healthcare datasets. By creating synthetic instances that represent underrepresented characteristics or conditions, this approach contributes to the development of more accurate and unbiased predictive models in medical research and decision-making.

However, despite the high potential of in-silico trials, we foresee some limitations both upstream and downstream of their activation. From a legal perspective, as of the latest European Union regulation, the use of synthetic data is addressed in the recently enacted AI Act (December 9, 2023), but specific national legislation is yet to be established. Consequently, the practical application of a synthetic cohort in a virtual trial could face resistance from national regulatory authorities unaccustomed to dealing with this matter.

From a patient perspective, potential reluctance to share anonymized data or skepticism toward participating in in-silico trials [17] is an important consideration.

Additionally, inherent limitations arise from synthetic data. When employing synthetic patients as a control cohort, we encounter similar challenges as those in non-RCT studies, particularly concerning confounders. Synthetic patient cohorts may not comprehensively account for unknown confounding variables, which are variables not measured or considered during the cohort creation process.

Moreover, as previously mentioned, when utilizing a synthetic cohort derived from a specific country, its applicability should generally be restricted to an in-silico trial conducted within the same context.

In conclusion, we provided evidence of the feasibility of creating a virtual cohort of patients that faithfully replicates the original population, offering numerous advantages for trial development. Although our focus has been on constructing data that are typical in population-based leukemia survival research, this approach can readily be extended to other time-to-event settings, requiring only a set of covariates and a preliminary understanding of prognostic covariates.

We strongly believe that the simplicity of these methods offers a straightforward and easily implementable technique for generating synthetic data that maintains high data utility standards without compromising the validity of results. This strategy first applied to AML, could be extended to other diseases to improve the prognosis and the management of clinical trials in other settings. Synthetic patient generation may ultimately simplify and accelerate the evaluation of new therapies and enable faster access to innovative treatments.

AUTHOR CONTRIBUTIONS

Conception and design: Alfonso Piciocchi, Marta Cipriani, Paola Fazi, and Marco Vignetti. Provision of study materials or patients: Giovanni Marconi, Giovanni Martinelli, Adriano Venditti, Maria Teresa Voso, Francesco Buccisano, Maria Valeria Feraco, and Marco Ferrante. Collection and assembly of data: Valentina Arena, Stefano Soddu, Enrico Crea, and Edoardo La Sala. Data analysis and interpretation: Alfonso Piciocchi, Marta Cipriani, and Monica Messina. Manuscript writing: Alfonso Piciocchi, Marta Cipriani, and Monica Messina. Final approval of manuscript: All authors. Accountable for all aspects of the work: Alfonso Piciocchi and Marta Cipriani.

CONFLICT OF INTEREST STATEMENT

- Giovanni Marconi acts as a consultant or is included in the speaker bureau of Abbvie, Astellas, AstraZeneca, Immunogen, Jansenn, Menarini/Stemline, Pfizer, Ryvu, Servier, Syros, Takeda and received research funding from Abbvie, Astellas, AstraZeneca, Daiichi Sankyo, and Pfizer.
- Adriano Venditti acts as a consultant or is included in the speaker bureau of Novartis, Pfizer, Jazz Pharmaceuticals, Amgen, AbbVie, Gilead, Astellas, Incyte, Medac, AstraZeneca, Servier, Daiichi-Sankyo, Janssen-Cilag, Menarini-StemLine Laboratories Delbert, and Bristol Myers Squibb and received research funding from Sandoz and Jazz Pharmaceuticals.

All other co-authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

The GIMEMA AML1310 study was approved by a Medical Ethics Committee. All procedures in this study were performed under the principles of the Declaration of Helsinki and the institutional guidelines.

PATIENT CONSENT STATEMENT

The authors have confirmed that written informed consent was obtained from the patients enrolled in the GIMEMA AML1310 study.

CLINICAL TRIAL REGISTRATION

The GIMEMA AML1310 trial was registered at www.clinicaltrials.gov as #NCT01452646 and EudraCT as #2010-023809-36.

ORCID

Marta Cipriani https://orcid.org/0000-0001-6159-8059
Monica Messina https://orcid.org/0000-0003-4078-7066
Adriano Venditti https://orcid.org/0000-0002-0245-0553

REFERENCES

1. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, et al. Application of a neural network whole transcriptome–based pan-cancer method for diagnosis of primary and metastatic cancers. JAMA Netw Open. 2019;2(4):e192597.
2. Khanani S. Editorial comment: artificial intelligence in mammography—our new reality. Am J Roentgenol. 2022;219(3): 381–381.
3. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019;10(1):3069.
4. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. 2014.
5. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton, FL: Routledge; 2017.
6. Little C, Elliot M, Allmendinger R, Samani SS. Generative adversarial networks for synthetic data generation: a comparative study. 2021.
7. Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: concepts and early adoptions. Brief Bioinform. 2019;20(5):1699–708.
8. Venditti A, Piciocchi A, Candoni A, Melillo L, Calafiore V, Cairoli R, et al. GIMEMA AML1310 trial of risk-adapted, MRD-directed therapy for young adults with newly diagnosed acute myeloid leukemia. Blood. 2019;134(12):935–45.
9. Buccisano F, Palmieri R, Piciocchi A, Arena V, Candoni A, Melillo L, et al. ELN2017 risk stratification improves outcome prediction when applied to the prospective GIMEMA AML1310 protocol. Blood Adv. 2022;6(8):2510–16.
10. Nowok B, Raab GM, Dibben C. synthpop: bespoke creation of synthetic data in *R*. J Stat Softw. 2016;74(11):1–26.
11. Elliot M. Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. 2014.
12. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. J R Stat Soc Ser A Stat Soc. 2018;181(3):663–88.
13. Raab GM, Nowok B, Dibben C. Assessing, visualizing and improving the utility of synthetic data. 2021.
14. D'Amico S, Dall'Olio D, Sala C, Dall'Olio L, Sauta E, Zampini M, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. JCO Clin Cancer Inform. 2023;7:e2300021.
15. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. 2021;11(4):e043497.
16. Rajotte JF, Bergen R, Buckeridge DL, El Emam K, Ng R, Strome E. Synthetic data as an enabler for machine learning applications in medicine. iScience. 2022;25(11):105331.
17. Ghafur S, Van Dael J, Leis M, Darzi A, Sheikh A. Public perceptions on data sharing: key insights from the UK and the USA. Lancet Digit Health. 2020;2(9):e444–46.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Piciocchi A, Cipriani M, Messina M, Marconi G, Arena V, Soddu S, et al. Unlocking the potential of synthetic patients for accelerating clinical trials: Results of the first GIMEMA experience on acute myeloid leukemia patients. eJHaem. 2024;5:353–59. https://doi.org/10.1002/jha2.873