ELSEVIER

Blood 146 (2025) 4350–4351

The 67th ASH Annual Meeting Abstracts

POSTER

803. EMERGING TOOLS, TECHNIQUES, AND ARTIFICIAL INTELLIGENCE IN HEMATOLOGY

Development and validation of synthetic data generation over a federated learning computing framework to accelerate innovation and boost personalized medicine in hematological diseases

Gianluca Asti[1], *Mattia Delleani*[2], *Patricia Apellaniz*[3], *Imanol Isasa*[4,5], *Daniela Martinez Duarte*[6], *Borja Arroyo Galende*[3], *Francesco Casadei*[7], *Eleonora Iascone*[1], *Marilena Bicchieri*[1], *Elisabetta Sauta*[1], *Elena Zazzetti*[1], *Alessia Campagna*[1], *Giulia Maggioni*[1], *Luca Lanino*[8], *Alessandro Buizza*[1], *Ivan Ferrari*[1], *Alessandro Bruseghini*[1], *Matteo Zampini*[1], *Alejandro Almodóvar*[3], *Silvia Uribe*[3], *Victor Savevski*[1], *Antonio Almeida*[9,10], *Rami Komrokji*[11], *Amer Zeidan*[12], *Pierre Fenaux*[13], *Lin-Pierre Zhao*[13], *Lars Bullinger*[14], *Sträng Eric*[14], *Maria Diez-Campelo*[15], *Juan Parras*[3], *Leonor Cerdá Alberich*[16], *Gastone Castellani*[17], *Santiago Zazo*[3], *Rudolf Mayer*[6], *Andoni Beristain*[4,18], *Federico Alvarez*[3], *Saverio D'Amico*[2], *Matteo Della Porta*[1,19]

[1] Humanitas Clinical and Research Center, IRCCS, Rozzano, Italy
[2] Train s.r.l., Rozzano, Italy
[3] Information Processing and Telecommunications Center, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
[4] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA);, Donostia-San Sebastián, Spain
[5] Computer Science and Artificial Intelligence Department, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain
[6] SBA Research, Vienna, Austria
[7] Bioinformatics Laboratory, IRCCS Institute of Neurological Sciences of Bologna, Bologna, Italy
[8] Yale University, New Haven, United States
[9] Hospital da Luz, Department of Hematology, Lisboa, Portugal
[10] Faculdade de Medicina, Universidade Católica Portuguesa, Lisboa, Portugal
[11] H. Lee Moffitt Cancer Center, Tampa, United States
[12] Yale School of Medicine, Yale University, Department of Internal Medicine, New Haven, United States
[13] Hôpital Saint-Louis, Université de Paris, Paris, France
[14] Charité – Universitätsmedizin Berlin, Berlin, Germany
[15] Hospital Clínico Universitario de Salamanca, Salamanca, Spain
[16] Grupo de Investigacion Biomedica en Imagen (GIBI230), Instituto de Investigacion Sanitaria La Fe, Valencia, Spain
[17] University of Bologna, Department of Medical and Surgical Sciences, Bologna, Italy
[18] eHealth Group, Biogipuzkoa Health Research Institute, Donostia-San Sebastián, Spain
[19] Humanitas University, Department of Biomedical Sciences, Pieve Emanuele, Italy

Abstract Background: Access to large, diverse, and high-quality datasets is essential for advancing research in rare hematological diseases (RHDs). However, privacy regulations and institutional silos often limit data sharing across centers. Synthetic data (SD) generation using deep learning offers a promising solution to augment clinical datasets while preserving patients confidentiality. Merging generative Artificial Intelligence (AI) training with federated learning (FL), which enables decentralized model training across institutions without data sharing, is particularly suited for healthcare applications. SD offer a series of transformative advantages for RHDs: by mimicking clinical and genomic characteristics of real patients without reproducing identifiable information, SD enable secure and scalable data sharing that respects privacy constraints. Furthermore, SD allow for data augmentation, missing data imputation and cohort balancement. SD have demonstrated high fidelity in replicating survival outcomes, molecular profiles, and complex gene-gene interactions observed in real populations. They have been successfully used to anticipate molecular classifications and prognostic scoring systems, demonstrating their potential to accelerate translational research. Furthermore, SD can support clinical trial innovation as synthetic control arms, reducing the need for placebo groups and streamlining trial design. Together, the integration of SD generation and FL offers a privacy-preserving, high-utility framework for advancing precision medicine and collaborative research in RHDs.

Aims: This project was conducted within the Synthema and Synthia Consortia. Specifically, the project aimed to 1) Implement and compare generative models within a FL framework to synthesize high-fidelity patient data while preserving privacy; 2) Validate the statistical, clinical, and privacy performance of generated SD using Synthetic vAlidation FramEwork (SAFE); 3) Assess the effectiveness of federated training strategies against centralized and isolated training scenarios to benchmark performance and ensure scalability in real-world clinical research networks.
Methods: A multi-institutional simulation was conducted using a myelodysplastic syndromes (MDS) dataset of 4427 patients distributed across three federated nodes. Several generative models (CTGAN, Bayesian Networks, and VAE-BGM) were trained under four FL strategies, including Federated Averaging and SD sharing. Data quality was assessed using the SAFE framework, which evaluates statistical fidelity, clinical relevance (including genomic associations and survival analysis), and privacy risk (via Nearest Neighbor Distance Ratio, NNDR). Model performance was compared against centralized and isolated training settings.
Results: Across the FL training rounds, SD demonstrated strong alignment with real-world data. Models trained via FL achieved high Clinical Synthetic Fidelity (CSF) and Genomic Synthetic Fidelity (GSF) scores, comparable to the centralized (upper-bound) setting and clearly superior to isolated node training. By training round 5, CSF and GSF reached 0.942 and 0.902 respectively, indicating a high degree of statistical and clinical similarity between real and SD. Privacy metrics, including the Nearest Neighbor Distance Ratio (NNDR), confirmed that the SD maintained strong privacy safeguards. Performance improvements were observed across rounds, as shown by the increasing fidelity scores, illustrating the learning benefits of model collaboration without data sharing. Furthermore, survival analysis and gene mutation frequency comparisons confirmed clinical utility of SD. SD preserved key genomic patterns and patient outcome distributions, validating their use in applications such as risk stratification and biomarker discovery.
Conclusions: This study demonstrates that FL can successfully generate high-fidelity SD while preserving patient privacy and data integrity. Generative AI trained with FL not only retain clinically relevant information but also achieve performance metrics comparable to centralized training, without requiring direct data sharing. These findings support the use of federated SD generation as a scalable and privacy-preserving solution for enabling secure multi-institutional research collaborations, particularly in settings where access to comprehensive medical data is limited, and for the development of personalized precision medicine AI models.

https://doi.org/10.1182/blood-2025-4350