# MOSAIC: An Artificial Intelligence–Based Framework for Multimodal Analysis, Classification, and Personalized Prognostic Assessment in Rare Cancers

Saverio D'Amico, MEng[1,2]; Lorenzo Dall'Olio, PhD[3]; Cesare Rollo, PhD[4]; Patricia Alonso, PhD[5]; Iñigo Prada-Luengo, PhD[6]; Daniele Dall'Olio, PhD[3]; Claudia Sala, PhD[7]; Elisabetta Sauta, PhD[1]; Gianluca Asti, MSc[1]; Luca Lanino, MD[1]; Giulia Maggioni, MD[1]; Alessia Campagna, MD[1]; Elena Zazzetti, MEng[1]; Mattia Delleani, MSc[1]; Maria Elena Bicchieri, PhD[1]; Pierandrea Morandini, MEng[1]; Victor Savevski, MEng[1]; Borja Arroyo, PhD[5]; Juan Parras, PhD[5]; Lin Pierre Zhao, MD[8]; Uwe Platzbecker, MD[9]; Maria Diez-Campelo, MD[10]; Valeria Santini, MD[11]; Pierre Fenaux, MD[8]; Torsten Haferlach, MD[12]; Anders Krogh, PhD[6]; Santiago Zazo, PhD[5]; Piero Fariselli, PhD[4]; Tiziana Sanavia, PhD[4]; Matteo Giovanni Della Porta, MD[1,13]; and Gastone Castellani, PhD[3,7]

DOI https://doi.org/10.1200/CCI.24.00008

## ABSTRACT

**PURPOSE** Rare cancers constitute over 20% of human neoplasms, often affecting patients with unmet medical needs. The development of effective classification and prognostication systems is crucial to improve the decision-making process and drive innovative treatment strategies. We have created and implemented MOSAIC, an artificial intelligence (AI)–based framework designed for multimodal analysis, classification, and personalized prognostic assessment in rare cancers. Clinical validation was performed on myelodysplastic syndrome (MDS), a rare hematologic cancer with clinical and genomic heterogeneities.

**METHODS** We analyzed 4,427 patients with MDS divided into training and validation cohorts. Deep learning methods were applied to integrate and impute clinical/genomic features. Clustering was performed by combining Uniform Manifold Approximation and Projection for Dimension Reduction + Hierarchical Density-Based Spatial Clustering of Applications with Noise (UMAP + HDBSCAN) methods, compared with the conventional Hierarchical Dirichlet Process (HDP). Linear and AI-based nonlinear approaches were compared for survival prediction. Explainable AI (Shapley Additive Explanations approach [SHAP]) and federated learning were used to improve the interpretation and the performance of the clinical models, integrating them into distributed infrastructure.

**RESULTS** UMAP + HDBSCAN clustering obtained a more granular patient stratification, achieving a higher average silhouette coefficient (0.16) with respect to HDP (0.01) and higher balanced accuracy in cluster classification by Random Forest (92.7% ± 1.3% and 85.8% ± 0.8%). AI methods for survival prediction outperform conventional statistical techniques and the reference prognostic tool for MDS. Nonlinear Gradient Boosting Survival stands in the internal (Concordance-Index [C-Index], 0.77; SD, 0.01) and external validation (C-Index, 0.74; SD, 0.02). SHAP analysis revealed that similar features drove patients' subgroups and outcomes in both training and validation cohorts. Federated implementation improved the accuracy of developed models.

**CONCLUSION** MOSAIC provides an explainable and robust framework to optimize classification and prognostic assessment of rare cancers. AI-based approaches demonstrated superior accuracy in capturing genomic similarities and providing individual prognostic information compared with conventional statistical methods. Its federated implementation ensures broad clinical application, guaranteeing high performance and data protection.

ACCOMPANYING CONTENT

Data Supplement

Accepted April 15, 2024
Published June 14, 2024

JCO Clin Cancer Inform
8:e2400008

## CONTEXT

**Key Objective**

Rare cancers constitute over 20% of human neoplasms, often affecting patients with unmet medical needs. Effective and standardized classification and prognostication systems on the basis of multimodal data analysis are crucial to improve the decision-making process and drive innovative treatment strategies.

**Knowledge Generated**

We developed and implemented MOSAIC, an artificial intelligence (AI)–based framework designed for multimodal analysis, classification, and personalized prognostic assessment in rare cancers. We compared different statistical and machine/deep learning–based approaches, using Explainable AI and federated learning for interpretability and distributed analysis.

**Relevance**

MOSAIC is a comprehensive AI-based framework to optimize classification and personalized prognostic assessment of rare cancers. The application of the model through a federated learning approach in clinical care can be valuable in decision-making process and driving treatment strategies.



## INTRODUCTION

According to RARECARENet, rare cancers are identified by an incidence of <six cases per 100,000 persons per year.[2] Twenty-five percent of patients diagnosed with cancer fall into this category, with each clinical entity within it encompassing a subset of patients with significant unmet medical needs. Rare cancers present unique challenges including low diagnostic rates, limited available data, and a lack of robust clinical evidence for treatment decisions.[3,4] Overall, these diseases constitute a public health emergency, underscoring the urgent need to devise new techniques for improved patient management.[5,6]

In this context, the establishment of effective classification and prognostication systems would offer immediate clinical utility, providing a solid basis to improve the clinical decision-making process. Conventional classification/prognostication tools in cancer primarily rely on clinical and histopathologic features, which are complemented by genomic features to better capture clinical-pathologic entities and predict clinical outcomes of interest.[7] The potential impact of genomic profiling in the classification and the clinical management of rare cancers lies on three key advantages: (1) enabling the categorization of morphologically defined neoplasms into distinct genomic subgroups with different therapeutic responses and outcomes, (2) identifying biomarkers for disease monitoring, and (3) laying the foundations for personalized treatments. In particular, the combined use of clinical and genomic data may enable the creation of prognostic models capable of generating personalized predictions of clinical outcomes.[8]

Hematologic neoplasms, such as myelodysplastic syndromes (MDSs), exemplify the clinical challenges associated with rare cancers.[9] MDSs are heterogeneous clonal hematopoietic disorders characterized by peripheral blood cytopenia and an increased risk of evolution into AML. These disorders range from indolent conditions to cases rapidly progressing into AML and, therefore, a risk-adapted treatment strategy is needed.[9,10] Over the past decade, genomic characterization has revealed a complex and heterogeneous landscape of recurrent genomic abnormalities in MDS, influencing distinct clinical phenotypes, survival rates, and disease progression risks.[11,12] Consequently, the assessment of mutational status has now been integrated into current classification and prognostication systems.[13,14]

Innovative technological approaches are essential to enhance the development of next-generation tools for personalized medicine in rare cancer and efficiently implement these models in clinical practice.[5,6] In this scenario, artificial intelligence (AI) holds great promise in health care.[15] Rare cancers, which are severely under-represented in basic and clinical research, can particularly benefit from AI technologies. The ability of AI technologies to integrate and analyze data from different sources can effectively address the unique challenges presented by rare cancers.[5,6]

In this study, we aimed to generate innovative, scalable, and fully explainable predictive models for patients with rare cancers combining statistical and machine learning methods. Moreover, we addressed the issue of implementing these models across multiple clinical centers, preserving data privacy. To this purpose: (1) we developed MOSAIC, an Artificial Intelligence–based Framework for Multimodal Analysis, Classification, and Personalized Prognostic Assessment in Rare cancers; (2) we proposed innovative algorithms to improve personalized medicine in these diseases; (3) we tested the framework's capability to implement clinical analysis in a specific use case (MDS), and (4) we provided preliminary evidence for the feasibility of a decentralized learning approach that enables collaborative

model training by distributing the learning process among several nodes, highlighting its advantages.

## METHODS

### Study Populations

The study was conducted by GenoMed4All[16] and Synthema[17] consortiums with the support of EuroBloodNET, the European Reference Network on rare hematologic diseases.[18] All the study procedures were compliant with the Ethics and governance of artificial intelligence for health.[19]

The selection criteria for the clinical use case were a cancer exhibiting rare prevalence in the general population with high clinical and molecular heterogeneities. This choice was aimed at assessing the performance of the MOSAIC platform under the most challenging clinical scenario. Consequently, we opted to focus on MDS as it aligned perfectly with these criteria.[9,11,12]

The study population consisted of a retrospective cohort of 2,043 patients provided by GenoMed4All and Synthema consortia (training cohort) and a publicly available data set of 2,384 patients by the International Working Group for the study of Prognosis in MDS (external validation cohort).[14] Inclusion criteria were age ≥18 years, diagnosis of MDS by WHO 2016 criteria,[20] and information available on demographics, clinical features, mutational screening/chromosomal abnormalities, treatment, and survival. The Humanitas Ethics Committee approved the study (ClinicalTrials.gov Identifier: NCT04889729). Written informed consent was obtained from each participant (Data Supplement, SF1).

### Design of MOSAIC Framework

MOSAIC, an Artificial Intelligence–based Framework for Multimodal Analysis, Classification, and Personalized Prognostic Assessment in Rare cancers, was designed to help clinicians in the implementation of next-generation classification and prognostic systems that integrate an increasing amount of genomic information (Fig 1). MOSAIC was developed with a strategic selection of methods, harnessing the full potential of multilayer data integration and analyses specific to rare diseases. We implemented advanced methods for real-world missing data imputation to prevent the introduction of potential biases that might compromise

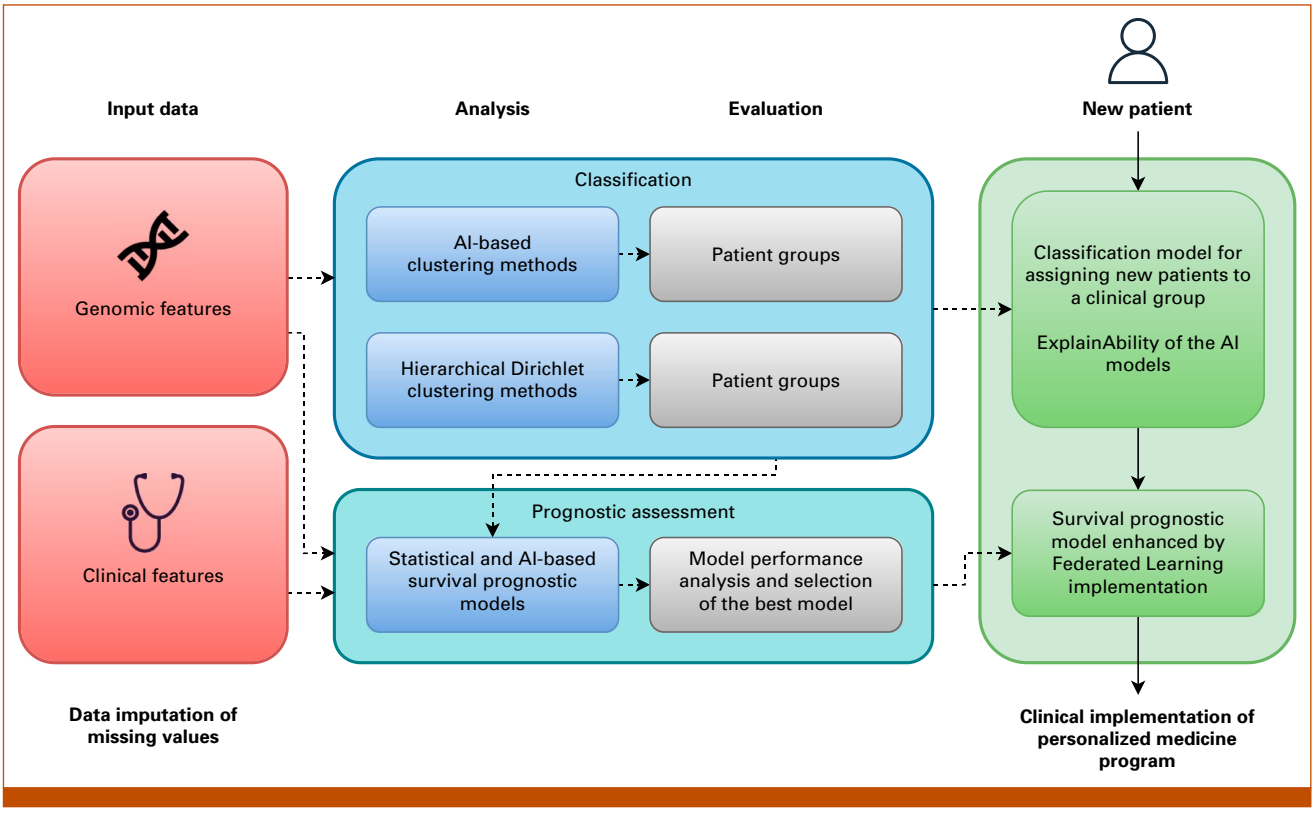



**FIG 1.** Overview of the MOSAIC framework architecture applied on training and validation cohorts. The figure shows the AI-based framework for multimodal analysis of classification and personalized prognostic assessment in rare cancers. Once the analysis framework is applied to the training cohort, the validated models can be used on new patients (green block), even in a federated environment. The scheme suggests analysis pathways, methods, and how to use them, for the multimodal analysis of classification and prognostic assessment in rare cancers, including the implementation of the models in a federated environment to enhance performance while maintaining a high degree of privacy. AI, artificial intelligence.

the generalizability of the predictive models.[21] High-performing hierarchical clustering methods were applied to genomic data to identify the main components defining the study population.[22] Information on the identified genomic-based patients' groups was then combined with demographics and longitudinal data to perform survival analysis, comparing statistical versus deep learning approaches. Models with the best performance were selected and included into a specific pipeline to allow the classification and prognostic assessment of new patients. In addition, MOSAIC enabled model implementation through a federated learning approach, which has provided evidence of improved prediction performance and a wide implementation of algorithms intended for clinical use, while maintaining a high degree of privacy.[23-27] The following sections show the results of the framework assessment we performed on patients with MDS. All the models were validated on the independent validation cohort.

### Missing Data Imputation Through Deep Learning

Deep generative decoder architecture was implemented for missing data imputation of the clinical-biologic variables (Data Supplement, Fig S1).[28] We used a latent space with a dimensionality of 2 and 10 Gaussian components. The neural network decoder contains two hidden layers with 12 and 33 units, respectively. We trained all parameters during 200 epochs using the Adam optimizer[29] (with betas 0.5 and 0.9) with a learning rate of 1e-1 for the Gaussian mixture model and representations and a learning rate of 1e-3 for the decoder parameters (Data Supplement, Fig S2). To perform imputation, we masked out 10% of the original MDS data set and tested imputation accuracy on the masked values. To yield the final imputed values, we left the trained Gaussian mixture model and decoder intact while maximizing the probability of the missing data conditioned on the observed data and a maximum-a-posteriori estimated representation (Data Supplement, SF2).

### Clinical and Genomic-Based Clustering and Classification of Rare Cancers

Hierarchical Dirichlet Process (HDP) was considered to define groups of patients according to specific genomic features.[12] In addition, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[22] was applied to the two-dimensional embedding of the genomic panel obtained from Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP),[30] which reproduces the data spatial distribution in a lower dimensional space. UMAP significantly reduced the number of features in the clustering step, avoiding the curse of dimensionality.[31] This is due to the capability of manifold learning algorithms (a subclass of dimensionality reduction) to create a low-dimensional embedding space that will be input to HDBSCAN, preserving the global information of the data spatial distribution. HDP identifies statistical components for molecular alterations, whereas UMAP + HDBSCAN focuses on clustering patients on the basis of the similarity of their molecular profiles. Both HDP and UMAP + HDBSCAN automatically tune the number of clusters. HDP is more robust than UMAP + HDBSCAN, but it may struggle with extreme heterogeneity and it requires slower computational times. On the other hand, UMAP + HDBSCAN is the state-of-the-art preferred approach because of its scalability on big data and efficient visualization capabilities, but it requires clinical feedback for optimal cluster determination, especially in heterogeneous and sparse data scenarios.[32] To improve UMAP stability and identify the optimal number of clusters, we tuned the number of the nearest neighbors by identifying the first significant gap in the average distance from the neighbors' trend (Data Supplement, Fig S3), which affects the cluster size. A statistical analysis was implemented to check if clusters found by UMAP + HDBSCAN may provide coherent and integrative information to HDP clusters. We measured the Adjusted Rand Index (ARI) for agreement among the two clustering schemes, and we implemented a Random Forest classification to assign new patients to a specific cluster in a clear, interpretable, and reproducible way to investigate cluster characterization and reproducibility. Models' explainability was performed through the Shapley Additive Explanations Approach (SHAP) to investigate features' importance and their effect on the cluster assignment process.[33] Further details are given in the Data Supplement (SF3).

### Prognostic Assessment of Rare Cancers on the Basis of the Integration of Clinical Parameters and Genomics

The following models were assessed to identify the best-performing prognostic tools for patients with rare cancers by integrating clinical and genomic features: Cox Proportional Hazards (CoxPH) model (and its penalized version),[34] Random Survival Forests,[35] DeepSurv,[36] Gradient Boosting,[37] and XGboost Survival.[38] All these models, except for CoxPH, are inherently nonlinear, thus potentially capable of leveraging complex interactions among the covariates to predict the survival outcome. We also considered different feature configurations, including patients' stratification from the previous step. Models' hyperparameters were optimized using a three-time repeated three-fold cross-validation schema on the training cohort, targeted to maximize the median Concordance-Index (C-Index). All models were then tested with a 10-time repeated three-fold cross-validation on the training cohort and further on the external validation cohort. Finally, we performed explainability analysis through SHAP on the outcome of the best model in terms of performance.[33] Further details are given in the Data Supplement (SF4).

### Federated Learning Implementation of the Survival Predictive Models

To provide evidence of the advantages provided by implementing the models using a federated approach, we simulated a federated environment in different centers (ie, hospitals providing data) to give an insight into how this

process would affect a clinical scenario by implementing three experimental settings. In setting A, we performed a centralized training of the CoxPH model on a single center (node) using all the 4,427 patients with MDS and simulating the best case scenario in which the model is trained on all the available data. In setting B, we assumed three nodes participating with a different amount of data for training the CoxPH (first node: 60% of the total number of patients with all features; second node: 30% of the total number of patients with all features; third node: 10% of the total number of patients, but forcing a significantly high percentage of missing covariates). This setting allows testing if a federated implementation helps to improve local results since it is known that the node with fewer data will perform worse than the other one. Finally, in setting C, a federated approach was implemented by using Federated Average (FedAvg)[39] on the three nodes of the previous setting. Every node independently trains the model and periodically transmits its model weights to a central model. The central model then computes an average of the weights, assigning greater significance to nodes with a higher patient count. This average is then transmitted back to each node to update its model weights (Data Supplement, SF5).

## RESULTS

### Missing Data Imputation

We applied an advanced deep generative decoder for data imputation,[28,29] achieving accurate results. We first focused on clinical-biologic features, obtaining low errors in imputing missing values. When imputing genomic features, we observed a tendency to assign higher mutation probabilities to mutated sites, indicating that the model is able to learn which sites are more likely to be mutated in MDS. Comparative analysis revealed the decoder's superiority over traditional mean imputation methods for both clinical and genomic features.

### Patient Classification on the Basis of Genomic Data

We performed patient classification using HDP and UMAP + HDBSCAN clustering approaches. We noticed that both approaches identified valid clusters in our training cohort, where HDP identified eight robust clusters, whereas UMAP + HDBSCAN identified 18 strongly characterized clusters, with a boost in granularity, but with considerable overlap with HDP outcomes (Fig 2A; Data Supplement, Fig S4), and 38.5% of ARI agreement, significantly good for a HDBSCAN clustering scheme with more than twice the number of HDP clusters. Indeed, the average Silhouette Coefficient on the same original data space was 0.16 for UMAP + HDBSCAN and 0.01 for HDP and, therefore, the higher data compression with UMAP + HDBSCAN might foster the creation of less heterogeneous groups by increasing the optimal number of clusters. Figure 2B shows how the patients assigned to HDP clusters are distributed in the UMAP + HDBSCAN clusters. Finally, a Random Forest classifier was used to assign patients from a test set to the clusters found, obtaining higher performance with UMAP + HDBSCAN with respect to HDP in terms of balanced accuracy (92.7% ± 1.3% *v* 85.8% ± 0.8%) and Cohen's K (92.1% ± 1.4% *v* 83.3% ± 0.9%) as shown in the Data Supplement (Table S1). Replicating these analyses on the validation cohort, similar distributions were observed when trying to assign patients from the external validation cohort to the UMAP + HDBSCAN clusters (Fig 3A) compared with the training cohort. The explainability analysis showed that, in both populations, similar features drive patients' classification (Fig 3B). Figure 3C shows SHAP summary plot analysis on the top 10 most important defining features for training and external validation cohorts for clusters 1, 2, 3, and 4 (others in the Data Supplement, Fig S5).

### Comparison of Prognostic Models for Rare Cancers

We applied different survival models considering demographics and clinical and genomic features to predict the probability of overall survival in patients with MDS for the training and the validation external cohort, which represented a blind test set to assess the performance.

The models' performance in the two cohorts (expressed in terms of C-Index) is reported in Figure 4A. The results obtained by using different feature sets are shown in the Data Supplement (Fig S6).

Overall, nonlinear machine learning/deep learning–based methods outperformed classical CoxPH-based approaches. All the models showed higher C-indices with respect to the reference prognostic tool for MDS (Revised International Prognostic Scoring System).[10] In particular, the Gradient Boosting Survival (GBS) model achieved a mean C-Index of 0.74 (SD, 0.02) in the external validation cohort, significantly greater than that in the linear CoxPH method (C-Index, 0.71; SD, 0.02). C-Indices in the internal validation (from training cohort) were 0.75 (SD, 0.01) and 0.77 (SD, 0.01) for CoxPH and GBS, respectively. Explainability analysis by SHAP showed similar feature importance ranking for both training and external validation cohorts (Fig 4B).

### Federated Learning Implementation of the Predictive Models

We implemented FedAvg[39] on the CoxPH model[34] considering overall survival as the end point and performed a three-fold cross-validation evaluating the C-Index for all the experimental settings described in the Methods section (Fig 5A). In setting A, the centralized model results in a C-Index of 0.74 ± 0.01. Results in setting B in terms of C-Index were 0.72 ± 0.03, 0.70 ± 0.002, and 0.54 ± 0.11 for the nodes having 60%, 30%, and 10% of all available

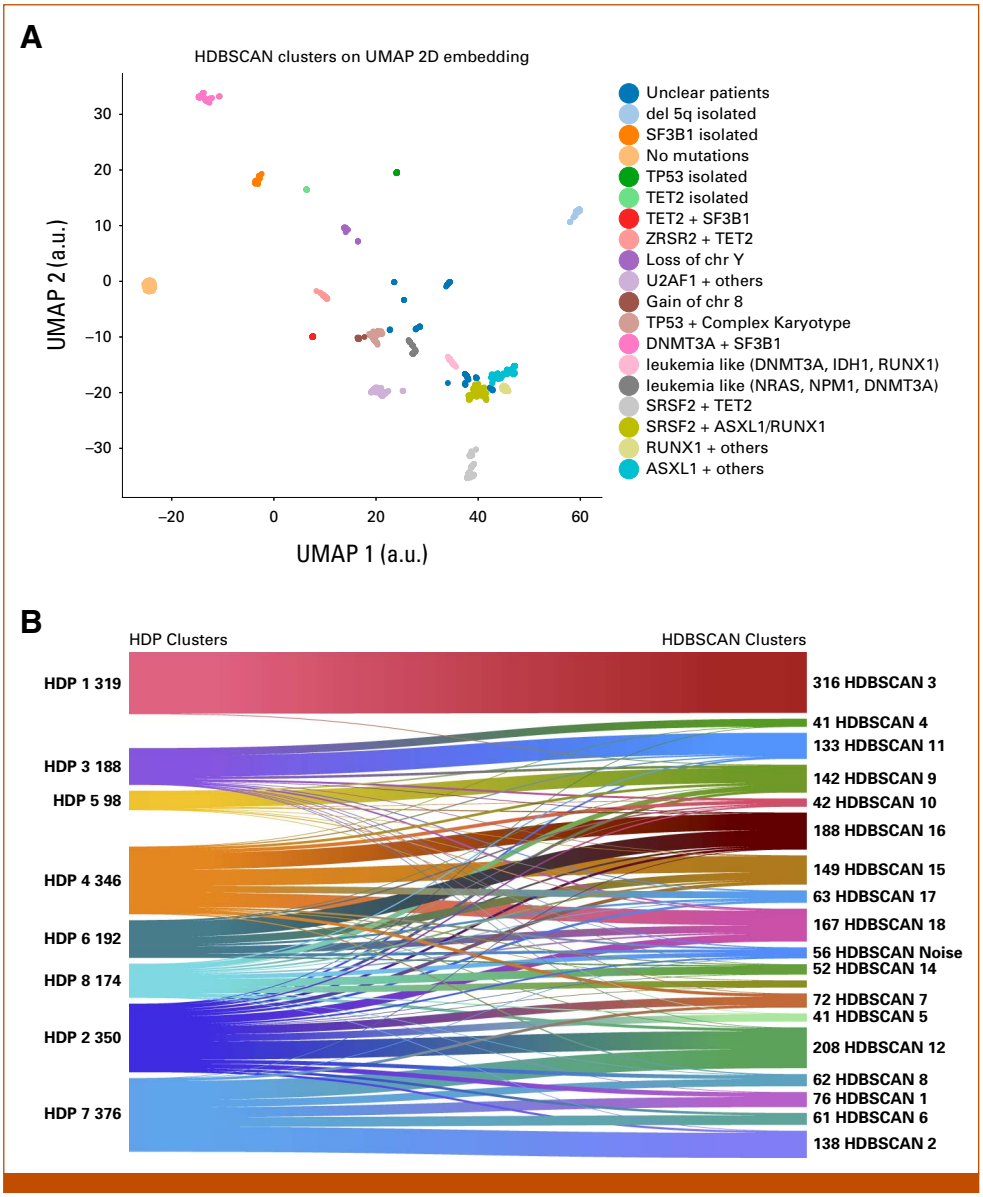



**FIG 2.** Patient clustering on the basis of genomic features performed using AI-based clustering and HDP methods on the MDS cohort (N = 2,043). (A) UMAP two-dimensional embedding. Each dot represents a patient, whose location is defined on the basis of its cytogenetics and genomic features (gene mutations). The figure shows the number of assigned clusters together with some labels to specify the genomic characterization of some clusters. The model found 18 clusters with 56 unclear patients assigned to cluster −1. (B) Alluvial plot showing the more granular classification of HDBSCAN compared with the clinical groups in the study by Bersanelli et al[12] found using the HDP clustering approach. AI, artificial intelligence; HDBSCAN, Hierarchical Density-Based Spatial Clustering of Applications with Noise; HDP, Hierarchical Dirichlet Process; MDS, myelodysplastic syndrome; UMAP, Uniform Manifold Approximation and Projection for Dimension Reduction.

patients, respectively. Finally, in setting C implementing the CoxPH model with FedAvg, results in terms of C-Index were 0.72 ± 0.02, 0.73 ± 0.01, and 0.63 ± 0.12 for the first, second, and third nodes, respectively. The obtained results summarized in Figures 5B and 5C demonstrate an enhancement in the model's performance with the implementation of the federated approach (Data Supplement, Fig S7), even in a situation of data scarcity (third node in settings B and C). Results for overall survival and event-free survival are reported in the Data Supplement (Fig S8 and Table S2).

## DISCUSSION

Next-generation classification/prognostication systems integrating clinical and morphologic information with the genomic profile are expected to improve the management of

patients with rare cancers.[4,5] Here, we showed that the MOSAIC platform offers an explainable and robust approach to optimize classification and prognostic evaluation of rare cancers, such as MDS. MOSAIC incorporates cutting-edge methodological approaches specifically designed to handle and integrate complex information, even in scenarios with a limited amount of data available (as reflected in federated experimental settings). Our findings indicate that machine/deep learning methods are more efficient and accurate than conventional statistical approaches in capturing genomic similarities between patients and prognostic information at the individual patient level. In addition, federated learning algorithms enable a wide clinical implementation of the models, ensuring high performance and data protection.

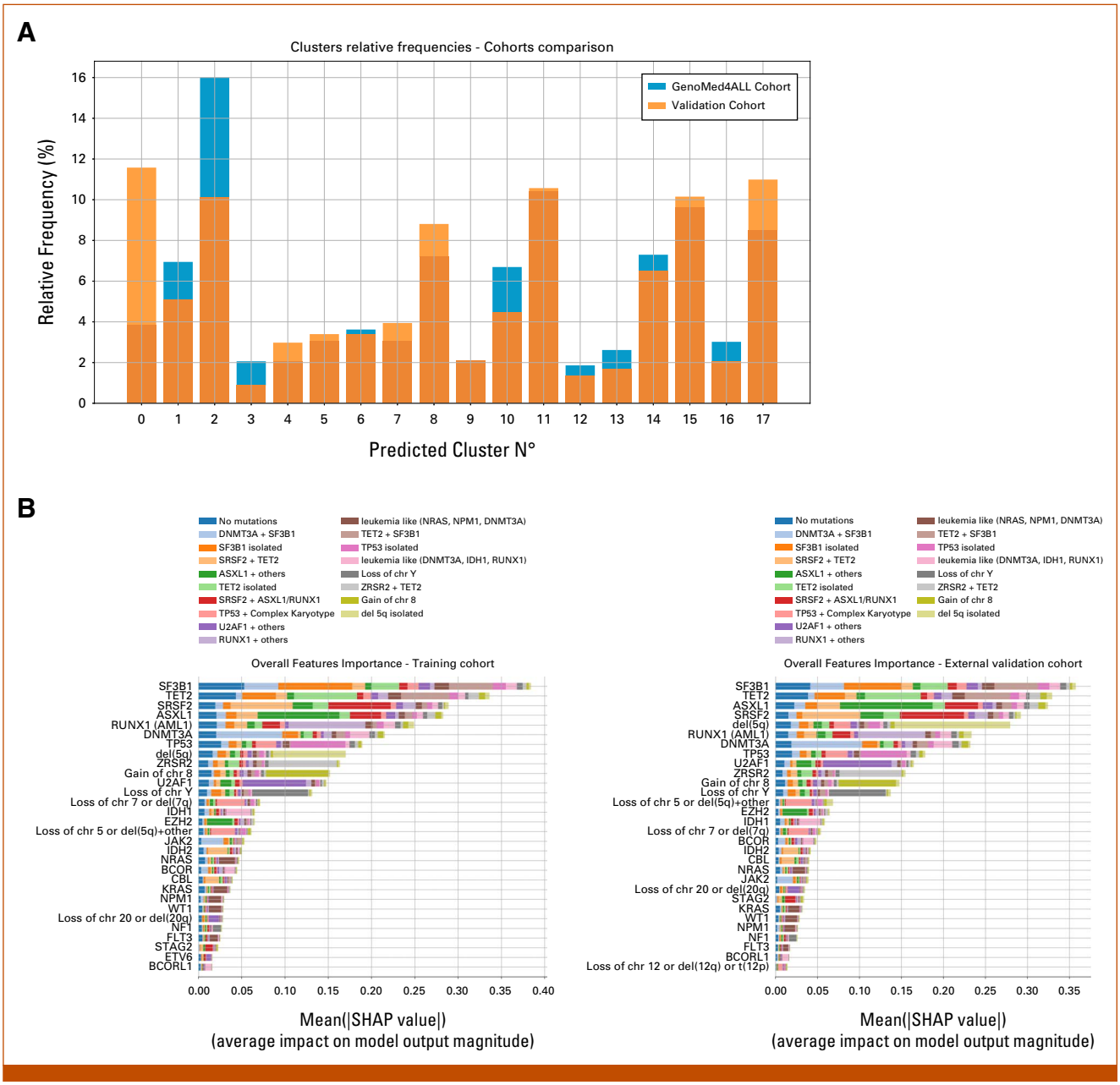



**FIG 3.** Validation of the identified clusters in the MDS cohorts (N = 4,427) using XAI frameworks. (A) Cluster relative frequency in both MDS training (N = 2,043) and validation cohorts (N = 2,384); clusters were assigned training a RF classifier (100 trees, maximum depth = 35, minutes samples per leaf = 1) on the whole training cohort. (B) Average impact on cluster assignment for every feature and every cluster, obtained using SHAP on the trained best selected RF classifier on the training data set (left). Average impact on cluster assignment for every feature and every cluster, obtained using SHAP on the trained best selected RF classifier on the validation data set (right). (C) Feature impact on cluster assignment, obtained using SHAP on the trained best selected RF classifier, for both training and validation cohorts. MDS, myelodysplastic syndrome; RF, Random Forest; SHAP, Shapley Additive Explanations Approach; XAI, Explainable artificial intelligence. (continued on following page)

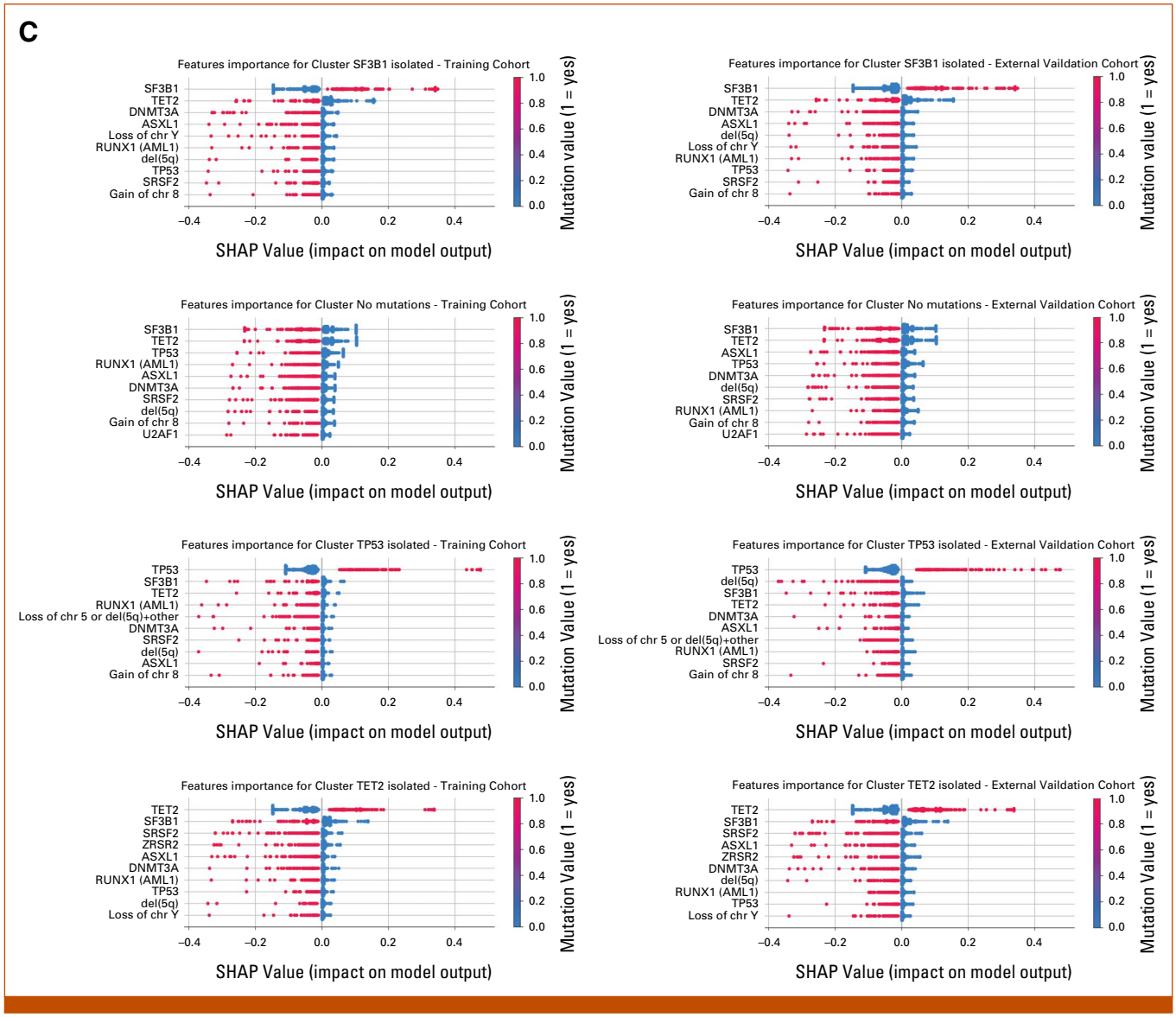



**FIG 3.** (Continued).

To assess the clinical value of the MOSAIC platform, different important issues were addressed, essential for the right deployment of predictive models in clinical care, particularly when including AI-based approaches. These include transparency, reliability, implementation, and data privacy preservability.[15,19]

In terms of transparency, MOSAIC provided a good understanding of all the models (interpretability and explainability) by using effective algorithms that assist clinicians in defining the most relevant clinical and genomic features (in order of priority) driving the classification and prediction of survival probabilities for an individual patient.

Regarding reliability, it is well known that main vulnerabilities of predictive models are related to lack of generalizability.[19] Here, we provided evidence for the clinical implementability of the MOSAIC platform in the challenging clinical scenario of a rare cancer with high clinical and genomic heterogeneities (MDS), also demonstrating the generalizability of findings through a large independent validation cohort.

Regarding technological implementability, the MOSAIC platform leverages containerized models, thus facilitating immediate deployment and continuous performance analysis in different hospitals, ready for federated infrastructure implementation.

A crucial topic for the clinical implementation of precision medicine tolls in rare cancer is the scalability of the technology across different clinical scenarios (rare cancers account for >20% of all human cancers, but each single identity involves only a limited number of patients and presents disease-specific clinical and genomic features).[2-6]

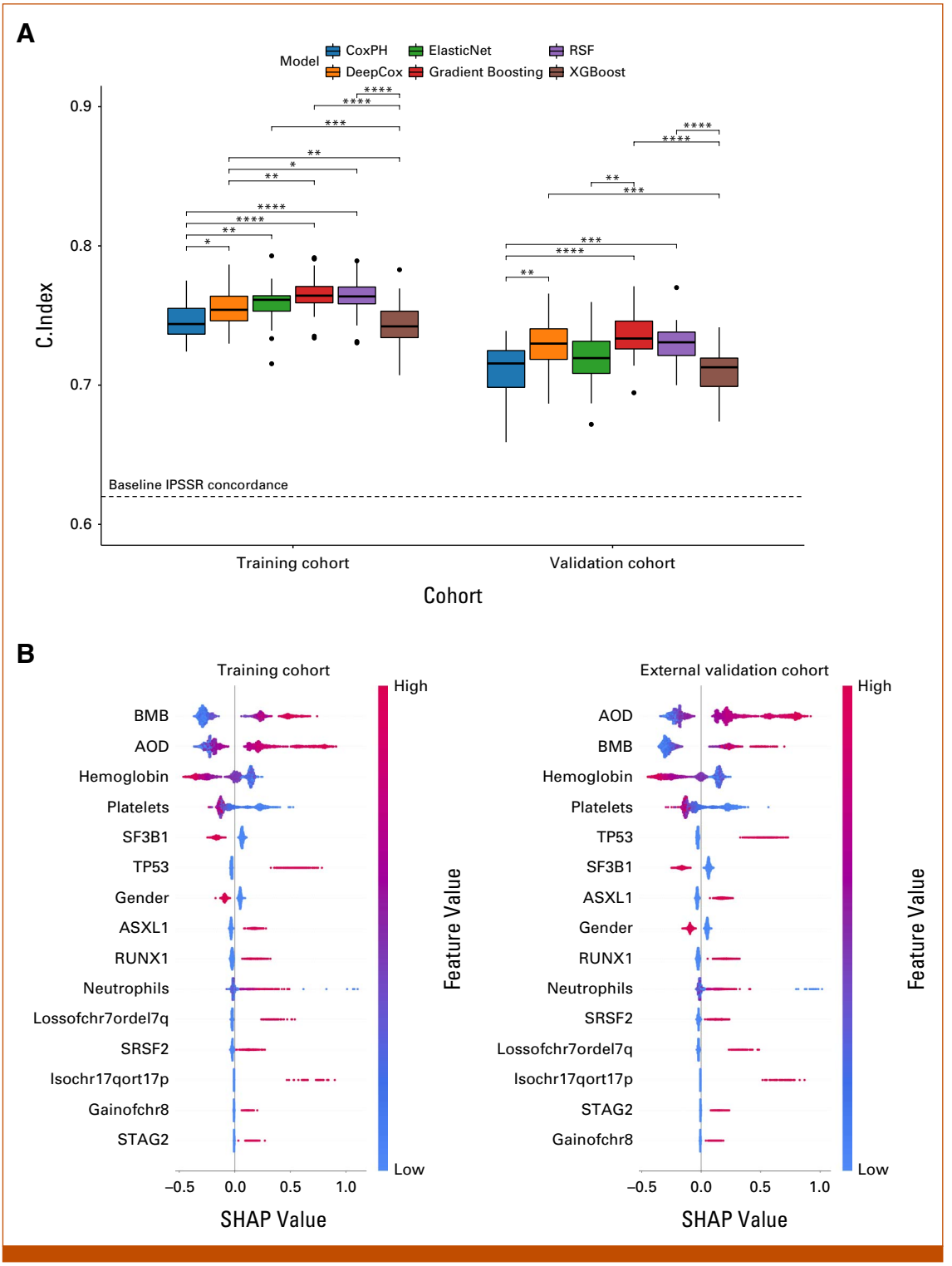



**FIG 4.** Prognostic assessment of patients with MDS (N = 2,043) on the basis of clinical and genomic features comparing different methods for survival prediction. (A) Comparison of different overall survival prediction methods in MDS: CoxPH model (and its penalized version), Random Survival Forests, DeepCox, Gradient Boosting, and XGboost survival methods. C-Index was used to evaluate model performance; C-Index of the conventional IPSS-R scoring system is reported as a baseline. *$P$ < .01, **$P$ < .001, ***$P$ < .0001. (B) Validation using XAI frameworks of the best-performing survival model (Gradient Boosting). The figure shows the features' impact on overall survival prediction in the training (right) and validation cohorts (left). C-Index, Concordance-Index; IPSS-R, Revised International Prognostic Scoring System; MDS, myelodysplastic syndrome; XAI, Explainable artificial intelligence.

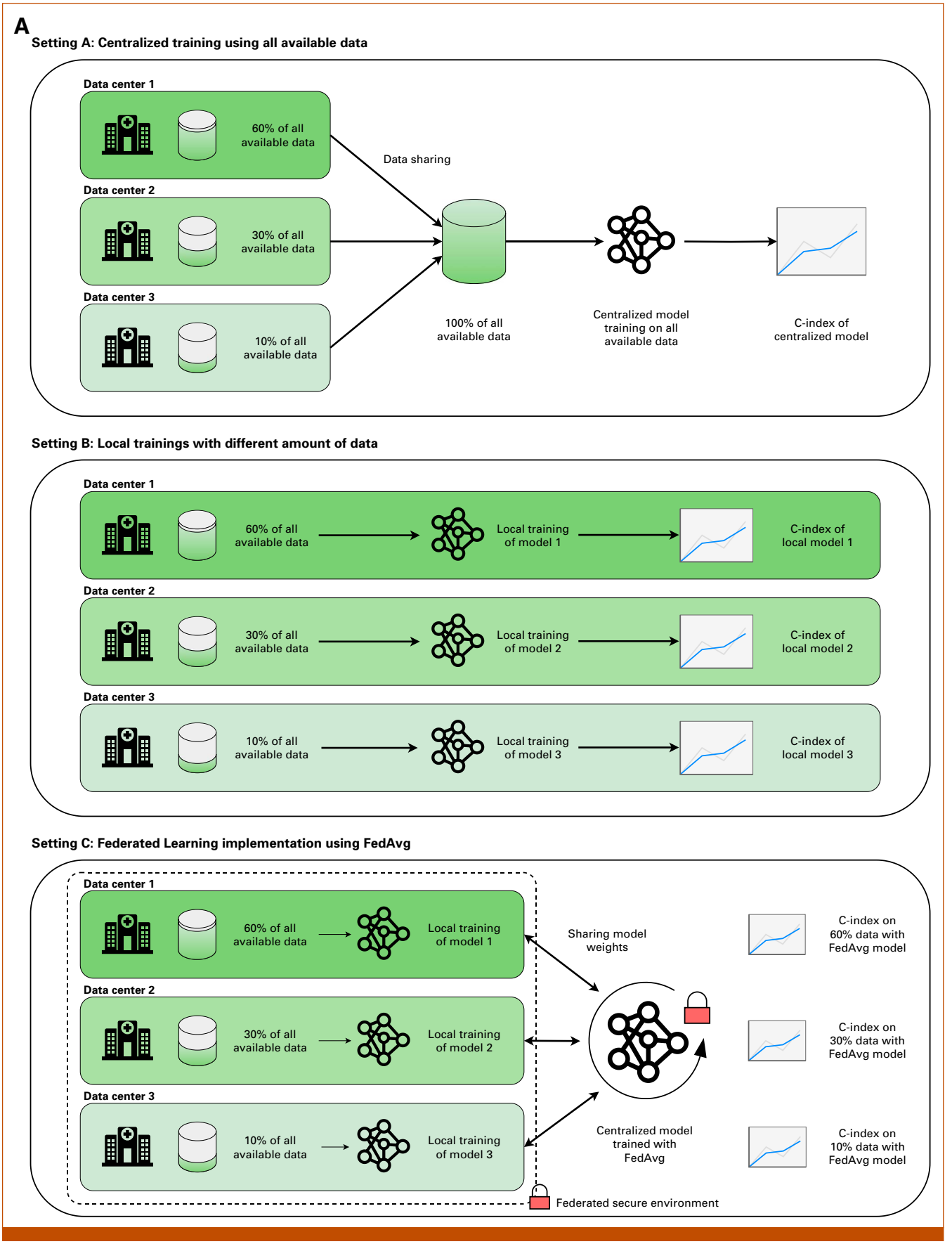
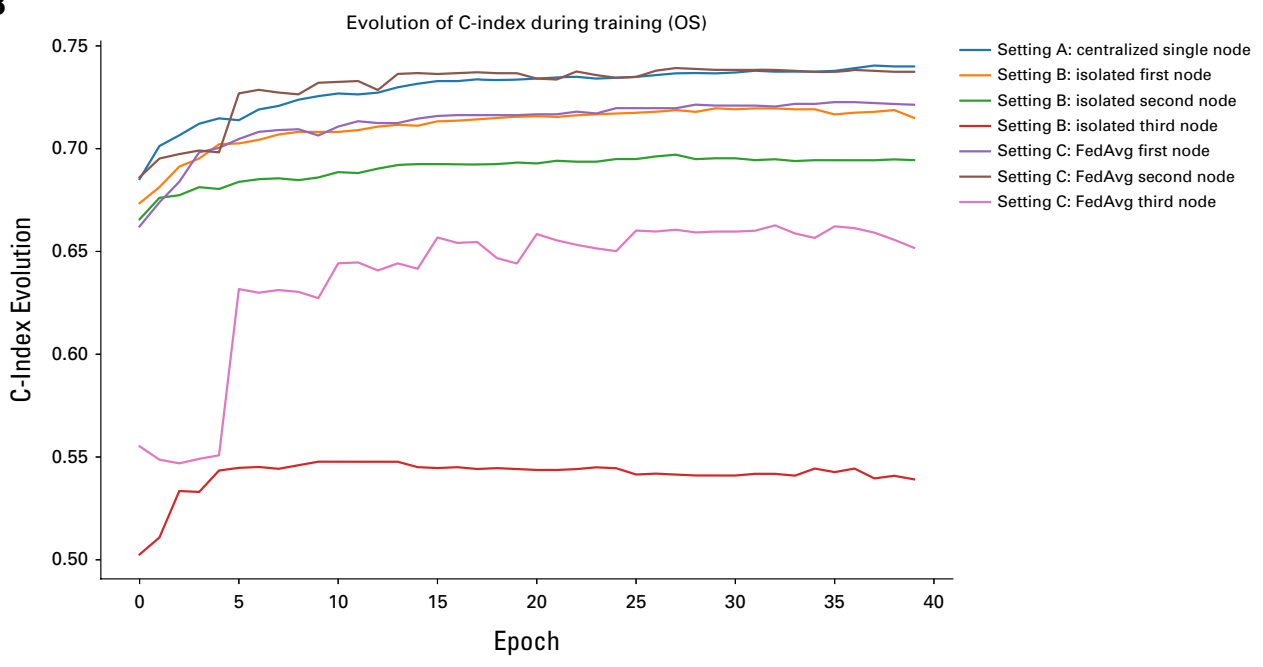



**FIG 5.** Federated learning implementation. (A) Overview of experimental settings implemented to test the benefits of a federated learning architecture. Setting C shows the federated architecture's implementation, allowing information of (continued on following page)




C

| Nodes | Overall Survival (C-Index) | | |
|---|---|---|---|
| | Cent | Isolated | FedAvg |
| Node 1 | 0.74 - 0.01 | **0.72 - 0.03** | **0.72 - 0.02** |
| Node 2 | | 0.7 - 0.002 | **0.73 - 0.01** |
| Node 3 | | 0.54 - 0.11 | **0.63 - 0.12** |

**FIG 5.** (Continued). individual models sharing without data transfer. In this setting, we simulated three different centers (ie, hospitals providing data) to have 60%, 30%, and 10% of the total MDS training patient population (N = 2,043). (B) The figure shows the evolution of the C-Index for overall survival calculated at each epoch during model training. It can be clearly observed how the value of this metric rises at five epochs in the nodes that train in a federated way. After this increase, they continue training with their data for another five epochs. This is why peaks can appear during training for each of the nodes, especially for nodes 2 and 3. (C) Experiment mean-SD results for C-Index metric. The results are evaluated for the overall survival. Bold entries refer to the best results for each node. C-Index, Concordance-Index; MDS, myelodysplastic syndrome; OS, overall survival; SD, standard deviation.



MOSAIC platform's ability to integrate multiple data layers and to account for conditional dependencies between features in a heterogeneous genomic landscape (MDS)[9-12] suggests that classification/prognostication models can be efficiently translated into different clinical scenarios. Multimodal analysis is crucial not only to increase the generalizability of models but also to increase the performances by leveraging more information. The next version of MOSAIC will involve the integration of other data modalities such as imaging, text, advanced omics data, and other nonstructured information available.

Innovative technologies for data collection and analysis to preserve data privacy are required for implementing personalized medicine, especially in rare cancers where predictive models rely on limited data. Federated learning addresses privacy concerns by collaboratively training algorithms without sharing data. Recent studies show that federated models perform comparably with centralized ones, without moving patient data beyond the firewalls of the institutions in which they reside. Our experiment demonstrates the effectiveness of federated learning in improving predictive models for MDS patient survival prediction.

MOSAIC aims to improve translational research by also scaling in different clinical domains of rare cancers. The initiative originates within the European GenoMed4All and Synthema consortia, following the goal of creating advanced, accessible, and clinically validated technologies in the clinical practice for precision medicine. For effective technology dissemination and validation, future plans include the following: (1) implementing the MOSAIC pipeline in all the reference centers of EuroBloodNET (>100) to increase the accuracy of classification/prognostication tools for rare hematologic diseases, (2) using the MOSAIC framework in validation studies on different rare cancers by comparing the

results with traditional methodologies (as done in MDS), and (3) designing and implementing dissemination and user adoption strategies to effectively raise awareness of results in different clinical settings by promoting the use of the MOSAIC open-source platform in other studies and implementation of new methodologies by the community.

## AFFILIATIONS

[1]Humanitas Clinical and Research Center—IRCCS, Milan, Italy
[2]Train s.r.l., Milan, Italy
[3]Department of Physics and Astronomy (DIFA), Bologna, Italy
[4]Computational Biomedicine Unit, Department of Medical Sciences, University of Turin, Turin, Italy
[5]Department of Signals, Systems and Radiocommunications, Polytechnic University of Madrid, Madrid, Spain
[6]University of Copenhagen, Copenhagen, Denmark
[7]Experimental, Diagnostic and Specialty Medicine—DIMES, Bologna, Italy
[8]Hematology and Bone Marrow Transplantation, Hôpital Saint-Louis/University Paris 7, Paris, France
[9]Medical Clinic and Policlinic 1, Hematology and Cellular Therapy, University Hospital Leipzig, Leipzig, Germany
[10]Hematology Department, Hospital Universitario de Salamanca, Salamanca, Spain
[11]Hematology, Azienda Ospedaliero-Universitaria Careggi & University of Florence, Florence, Italy
[12]MLL Munich Leukemia Laboratory, Munich, Germany
[13]Department of Biomedical Sciences, Humanitas University, Milan, Italy

## CORRESPONDING AUTHOR

Saverio D'Amico, MEng; e-mail: saverio.damico@humanitas.it.

## EQUAL CONTRIBUTION

S.D., L.D., C.R., P.A., and I.P.-L. contributed equally as first authors. M.G.D.P. and G.C. contributed equally as last senior authors.

## PRIOR PRESENTATION

Presented in part at the 64th American Society of Hematology (ASH) Annual Meeting, New Orleans, LA, December 10-13 2022.

## SUPPORT

Supported in part by European Union—Horizon 2020/2023 program (GenoMed4All project 101017549 to M.G.D.P.; U.P.; M.D-C.; P.F.; T.H.; G.C.; Synthema project, 1101095530 to M.G.D.P.; G.C.); Transcan 7 Horizon 2020 program (EuroMDS project 20180424 to M.G.D.P.; U.P.; P.F.); European Union—Next Generation EU—NRRP M6C2 (Investment 2.1 Enhancement and strengthening of biomedical research in the NHS—Project PNRR-MAD-2022-12376695 to M.G.D.P.); AIRC Foundation (Associazione Italiana per la Ricerca contro il Cancro, Milan Italy—Project 22053 and 29483 to M.G.D.P.; 26216 to G.C.; 26537 to V.S.); 5 × 1000 MYNERVA (Myeloid Neoplasms Research Venture AIRC—project 21267 to M.G.D.P.); PRIN (Ministry of University & Research, Italy—Project 20229B28PE and Project 2017WXR7ZT to M.G.D.P. and G.C.); Ricerca Finalizzata 2018 (Italian Ministry of Health, Italy—Project NET-2018-12365935 to M.G.D.P.). A.K. was funded by two grants from the Novo Nordisk Foundation: Center for Basic Machine Learning Research in Life Science (NNF20OC0062606) and Quantum for Life (NNF20OC0059939).

## DATA SHARING STATEMENT

If the work will eventually be accepted and published, public access will be provided to the official GenoMed4All/Synthema GitHub repositories containing data and source code implemented within the study.

## AUTHOR CONTRIBUTIONS

**Conception and design:** Saverio D'Amico, Cesare Rollo, Iñigo Prada-Luengo, Gianluca Asti, Elena Zazzetti, Mattia Delleani, Maria Elena Bicchieri, Victor Savevski, Borja Arroyo, Pierre Fenaux, Santiago Zazo, Piero Fariselli, Matteo Giovanni Della Porta, Gastone Castellani
**Financial support:** Matteo Giovanni Della Porta, Santiago Zazo, Gastone Castellani
**Administrative support:** Borja Arroyo
**Provision of study materials or patients:** Uwe Platzbecker, Maria Diez-Campelo, Pierre Fenaux
**Collection and assembly of data:** Saverio D'Amico, Elisabetta Sauta, Gianluca Asti, Luca Lanino, Giulia Maggioni, Alessia Campagna, Elena Zazzetti, Mattia Delleani, Lin Pierre Zhao, Uwe Platzbecker, Valeria Santini, Pierre Fenaux, Torsten Haferlach, Matteo Giovanni Della Porta
**Data analysis and interpretation:** Saverio D'Amico, Cesare Rollo, Patricia Alonso, Iñigo Prada-Luengo, Claudia Sala, Elisabetta Sauta, Gianluca Asti, Luca Lanino, Elena Zazzetti, Mattia Delleani, Pierandrea Morandini, Juan Parras, Uwe Platzbecker, Maria Diez-Campelo, Pierre Fenaux, Torsten Haferlach, Anders Krogh, Tiziana Sanavia, Matteo Giovanni Della Porta
**Manuscript writing:** All authors
**Final approval of manuscript:** All authors
**Accountable for all aspects of the work:** All authors

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

**Iñigo Prada-Luengo**
**Stock and Other Ownership Interests:** BioNano Genomics
**Travel, Accommodations, Expenses:** Euroimmun

**Uwe Platzbecker**
**Honoraria:** Celgene/Jazz, AbbVie, Curis, Geron, Janssen
**Consulting or Advisory Role:** Celgene/Jazz, Novartis, BMS GmbH & Co KG
**Research Funding:** Amgen (Inst), Janssen (Inst), Novartis (Inst), BerGenBio (Inst), Celgene (Inst), Curis (Inst)
**Patents, Royalties, Other Intellectual Property:** Part of a patent for a TFR-2 antibody (Rauner et al. Nature Metabolics 2019)
**Travel, Accommodations, Expenses:** Celgene

**Maria Diez-Campelo**
**Honoraria:** Celgene, Novartis, Keros Therapeutics
**Consulting or Advisory Role:** Celgene, Novartis, GlaxoSmithKline, Blueprint Medicines, Agios, Hemavan, Syros, Keros Therapeutics, Curis
**Travel, Accommodations, Expenses:** Gilead Sciences

**Valeria Santini**
**Honoraria:** Celgene/Bristol Myers Squibb, Novartis
**Consulting or Advisory Role:** Celgene/Bristol Myers Squibb, Novartis, Gilead Sciences, AbbVie, Syros Pharmaceuticals, Servier, Geron, CTI, Otsuka, Curis
**Research Funding:** Celgene (Inst)
**Travel, Accommodations, Expenses:** Janssen-Cilag, Celgene

**Pierre Fenaux**
**Honoraria:** Bristol Myers Squibb
**Consulting or Advisory Role:** Bristol Myers Squibb
**Research Funding:** Bristol Myers Squibb

**Torsten Haferlach**
**Employment:** MLL Munich Leukemia Laboratory
**Leadership:** MLL Munich Leukemia Laboratory

**Anders Krogh**
**Company:** AJ Vaccines (I)

No other potential conflicts of interest were reported.

## REFERENCES


1. Reference deleted
2. RARECARENet (information network on rare cancers). https://www.rarecarenet.eu/
3. Gatta G, van der Zwan JM, Casali PG, et al: Rare cancers are not so rare: The rare cancer burden in Europe. Eur J Cancer 47:2493-2511, 2011
4. DeSantis CE, Kramer JL, Jemal A: The burden of rare cancers in the United States. CA Cancer J Clin 67:261-272, 2017
5. Billingham L, Malottki K, Steven N: Research methods to change clinical practice for patients with rare cancers. Lancet Oncol 17:e70-e80, 2016
6. Blay JY, Coindre JM, Ducimetière F, et al: The value of research collaborations and consortia in rare cancers. Lancet Oncol 17:e62-e69, 2016
7. WHO classification of tumors. https://whobluebooks.iarc.fr/
8. Cazzola M, Sehn LH: Developing a classification of hematologic neoplasms in the era of precision medicine. Blood 140:1193-1199, 2022
9. Cazzola M: Myelodysplastic syndromes. N Engl J Med 383:1358-1374, 2020
10. Greenberg PL, Tuechler H, Schanz J, et al: Revised international prognostic scoring system for myelodysplastic syndromes. Blood 120:2454-2465, 2012
11. Cazzola M, Della Porta MG, Malcovati L: The genetic basis of myelodysplasia and its clinical relevance. Blood 122:4021-4034, 2013
12. Bersanelli M, Travaglino E, Meggendorfer M, et al: Classification and personalized prognostic assessment on the basis of clinical and genomic features in myelodysplastic syndromes. J Clin Oncol 39:1223-1233, 2021
13. Khoury JD, Solary E, Abla O, et al: The 5th edition of the World Health Organization classification of haematolymphoid tumours: Myeloid and histiocytic/dendritic neoplasms. Leukemia 36: 1703-1719, 2022
14. Bernard E, Tuechler H, Greenberg PL, et al: Molecular international prognostic scoring system for myelodysplastic syndromes. NEJM Evid 1:EVIDoa2200008, 2022
15. Rajpurkar P, Chen E, Banerjee O, et al: AI in health and medicine. Nat Med 28:31-38, 2022
16. GenoMed4All: Genomics for next generation healthcare. www.genomed4all.eu
17. Synthema: Synthetic Haematological Data. www.synthema.eu
18. EuroBloodNet: European reference network for rare hematological diseases. www.eurobloodnet.eu
19. The World Health Organization: 2021 WHO guidance on ethics and governance of artificial intelligence for health. https://www.who.int/publications/i/item/9789240029200
20. Arber DA, Orazi A, Hasserjian R, et al: The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 127:2391-2405, 2016
21. Steyerberg EW, van Veen M: Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 60:979, 2007
22. Campello RJGB, Moulavi D, Sander J: Density-based clustering based on hierarchical density estimates, in Pei J, Tseng VS, Cao L, et al (eds): Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, Volume 7819. Berlin, Heidelberg, Springer, 2013, pp 160-172
23. Sheller MJ, Edwards B, Reina GA, et al: Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Sci Rep 10:12598, 2020
24. Crowson MG, Moukheiber D, Arévalo AR, et al: A systematic review of federated learning applications for biomedical data. PLoS Digit Health 1:e0000033, 2022
25. Rieke N, Hancox J, Li W, et al: The future of digital health with federated learning. NPJ Digit Med 3:119, 2020
26. Cremonesi F, Planat V, Kalokyri V, et al: The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform. J Biomed Inform 141:104338, 2023
27. Rollo C, Pancotti C, Birolo G, et al: SYNDSURV: A simple framework for survival analysis with data distributed across multiple institutions. Comput Biol Med 172:108288, 2024
28. Schuster V, Krogh A: The deep generative decoder: MAP estimation of representations improves modelling of single-cell RNA data. Bioinformatics 39:btad497, 2023
29. Kingma DP, Ba J: Adam: A method for stochastic optimization. ICLR: International Conference on Learning Representations 1-15, 2015
30. McInnes L, Healy J, Saul N, et al: UMAP: Uniform manifold approximation and projection. J Open Source Softw 3:861, 2018
31. Belkin M, Niyogi P: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15:1373-1396, 2003
32. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, et al: UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genet 15:e1008432, 2019
33. Lundberg SM, Lee SI: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30, 2017
34. Simon N, Friedman J, Hastie T, et al: Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 39:1-13, 2011
35. Ishwaran H, Kogalur UB, Blackstone EH, et al: Random survival forests. Ann Appl Stat 2:841-860, 2008
36. Katzman JL, Shaham U, Cloninger A, et al: DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18:24, 2018
37. Hothorn T, Bühlmann P, Dudoit S, et al: Survival ensembles. Biostatistics 7:355-373, 2006
38. Barnwal A, Cho H, Hocking T: Survival regression with accelerated failure time model in XGBoost. J Comput Graph Stat 31:1292-1302, 2022
39. McMahan B, Moore E, Ramage D, et al: Communication-efficient learning of deep networks from decentralized data. PMLR 54:1273-1282, 2017