BRIEF RESEARCH REPORT
published: 29 March 2021
doi: 10.3389/fonc.2021.657191

Personalized Survival Prediction of Patients With Acute Myeloblastic Leukemia Using Gene Expression Profiling

Adrián Mosquera Orgueira[1,2,3*†], *Andrés Peleteiro Raíndo*[1,2,3†], *Miguel Cid López*[1,2,3], *José Ángel Díaz Arias*[1,2], *Marta Sonia González Pérez*[1,2], *Beatriz Antelo Rodríguez*[1,2,3], *Natalia Alonso Vence*[1,2,3], *Laura Bao Pérez*[1,2,3], *Roi Ferreiro Ferro*[1,2], *Manuel Albors Ferreiro*[1,2], *Aitor Abuín Blanco*[1,2], *Emilia Fontanes Trabazo*[1,2], *Claudio Cerchione*[4], *Giovanni Martinnelli*[4], *Pau Montesinos Fernández*[5], *Manuel Mateo Pérez Encinas*[1,2,3] *and José Luis Bello López*[1,2,3]

[1] University Hospital of Santiago de Compostela (SERGAS), Department of Hematology, Santiago de Compostela, Spain, [2] Health Research Institute of Santiago de Compostela, Grupo de Investigación en Síndromes Linfoproliferativos, Santiago de Compostela, Spain, [3] Departamento de Medicina, University of Santiago de Compostela, Santiago de Compostela, Spain, [4] Hematology Unit, Istituto Tumori della Romagna IRST IRCCS, Meldola, Italy, [5] Hospital Universitari i Politècnic La Fe, Department of Hematology, Valencia, Spain

OPEN ACCESS

Edited by:
Alison M. Michie,
University of Glasgow,
United Kingdom

Reviewed by:
Eduardo Rego,
São Paulo State University, Brazil
Jatinder K. Lamba,
University of Florida, United States

***Correspondence:**
Adrián Mosquera Orgueira
adrian.mosquera@live.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Hematologic Malignancies, a section of the journal Frontiers in Oncology

Received: 22 January 2021
Accepted: 12 March 2021
Published: 29 March 2021

Citation:
Mosquera Orgueira A, Peleteiro Raíndo A, Cid López M, Díaz Arias JÁ, González Pérez MS, Antelo Rodríguez B, Alonso Vence N, Bao Pérez L, Ferreiro Ferro R, Albors Ferreiro M, Abuín Blanco A, Fontanes Trabazo E, Cerchione C, Martinnelli G, Montesinos Fernández P, Mateo Pérez Encinas M and Luis Bello López J (2021) Personalized Survival Prediction of Patients With Acute Myeloblastic Leukemia Using Gene Expression Profiling. Front. Oncol. 11:657191. doi: 10.3389/fonc.2021.657191

Acute Myeloid Leukemia (AML) is a heterogeneous neoplasm characterized by cytogenetic and molecular alterations that drive patient prognosis. Currently established risk stratification guidelines show a moderate predictive accuracy, and newer tools that integrate multiple molecular variables have proven to provide better results. In this report, we aimed to create a new machine learning model of AML survival using gene expression data. We used gene expression data from two publicly available cohorts in order to create and validate a random forest predictor of survival, which we named ST-123. The most important variables in the model were age and the expression of *KDM5B* and *LAPTM4B*, two genes previously associated with the biology and prognostication of myeloid neoplasms. This classifier achieved high concordance indexes in the training and validation sets (0.7228 and 0.6988, respectively), and predictions were particularly accurate in patients at the highest risk of death. Additionally, ST-123 provided significant prognostic improvements in patients with high-risk mutations. Our results indicate that survival of patients with AML can be predicted to a great extent by applying machine learning tools to transcriptomic data, and that such predictions are particularly precise among patients with high-risk mutations.

Keywords: acute myeloid leukemia, cancer, survival, machine learning, gene expression, prognosis

INTRODUCTION

Acute Myeloid Leukemia (AML) is a heterogeneous neoplasm characterized by cytogenetic and molecular alterations that drive patient prognosis. Currently established AML risk stratification guidelines, like the *European Leukemia Net* (ELN) risk classification (1), are based primarily on a limited number of cytogenetic and molecular variables. However, these guidelines don't take into account the whole mutational profile of AML, the different layers of biological complexity in the tumor and the complex intertwining between patient outcomes and the complexity of molecular interactions. Therefore, there is substantial room for improving predictions of AML survival. In this report we present a new machine learning model of AML patient survival based on gene expression data, which achieves high predictability independently of high risk mutations.

In the last years, the emergence of artificial intelligence has brought new expectations to the field of medicine, particularly for disease diagnosis and prognostication. Machine learning (ML) is a field of artificial intelligence that performs outcome prediction based on complex interactions between multiple variables by making little assumptions about the relationship between the dependent and independent variables (2). In ML, a model is trained with examples and not programmed with human-made rules (3). The implementation of ML-based survival models is becoming increasingly popular in order to provide patient-centered risk information that can assist both the clinician and the patient.

Survival prediction of AML patients has been extensively improved in the last decades. Several biomarker panels based on next-generation sequencing of multiple recurrently mutated or aberrantly expressed genes have been proposed to facilitate improved prognostic stratification. Although multiple somatic alterations have been associated with patient outcome, such as those in *NPM1*, *CEBPA*, *FLT3*, *IDH1*, *IDH2*, *KIT*, *WT1* and *RUNX1* (4), only mutations a few mutations are broadly employed in current clinical routine (5). Multiple studies have proposed novel biomarker panels and personalized survival prediction models aimed to improve AML prognostication. Sherve et al. (6) developed a novel prognostic model that incorporates clinical, cytogenetic and mutational data to determine personalized outcomes for each particular patient (6). In the same line, Patkar et al. (7) created a scoring model that provides a mechanism to risk stratify AML patients with mutated *NPM1* (7); whereas Gerstung et al. (8) reported statistical models that can generate personally tailored clinical decision support from all of the available prognostic information arising from a knowledge bank of AML cases (8). These studies evidence that the application of ML to clinical and molecular data has the potential to predict patient outcomes since the moment of diagnosis, and therefore it may help to improve therapeutic strategies in the field of AML.

In the present study, we applied ML algorithms to gene expression data from AML cases in order to create new individualized models of survival based on retrospective data, and to understand their relationship with high-risk mutations.

METHODS

Two databases available in the *Gene Expression Omnibus* were used for model training and validation. The GSE37642 database was used for training, as it contains gene expression data from 562 adult patients diagnosed with AML who were treated in the multicenter phase III AMLCG-1999 trial. Median age was 45 years (range 18-85 years, 32% aged ≥ 65). The GSE68833 database was used for validation, and it contains gene expression data from 137 adult patients with AML included in *The Cancer Genome Atlas* (TCGA) cohort, with a median age of 59 years (range 18-88 years, 33% aged ≥ 65). Both databases used partially overlapping microarray chips, so that we rank-normalized the log2-transformed expression estimates and selected a set of 44,366 common gene expression probes between both cohorts.

Unsupervised Gene Expression Clusterization

Briefly, the *Mclust* algorithm (9) was used in order to detect the 2 most likely clusters of patients according to the expression of each probe (*Mclust* function, parameter G = 2). Briefly, the *Mclust* algorithm determines the most likely set of clusters according to geometric properties (distribution, volume, and shape). An expectation-maximization algorithm is used for maximum likelihood estimation, and the best model is selected according to Bayes information criteria. The association of each of these probe-level clusters with overall survival was calculated using cox regression. Thereafter, those probes whose clusterization was significantly associated with survival (Bonferroni adjusted *p*-value < 0.05) were selected for multivariate clusterization using the same *Mclust* algorithm. Cluster prediction was performed on the test set using parameters estimated in the training cohort, and cox regression was used to verify the association of this clusterization with overall survival. The Shoenfeld's test was used to assess the proportional hazards assumption.

Survival Analysis

We analyzed gene expression association with overall survival using cox regression implemented in R. Assumption of proportional hazards was checked with Schoenfeld's method.

Random Forest Survival Analysis

We started our analysis by testing the association of each probe with overall survival in the training set using multivariate cox regression. Schoenfeld's method was used to assess the proportional hazards assumption.

Random forest survival models were created with the *rfsrc* function implemented in the *randomForestSRC* package in R (10). We decided to use this type of model because, in contrast with deep networks, random forest can quantify the relative importance of each variable, and thus enable the filtering of low-importance variables for model reduction and performance improvement. Parameter tuning was performed using the *tune.rfscr* function, which optimizes the *mtry* and nnodes variables. Random forests were implemented on survival data of the training cohort. Bootstrapping without replacement was performed with the default *by.node* protocol. Continuous rank probability score (CRPS) was calculated as the integrated Brier score divided by time, and represents the average squared distances between the observed survival status and the predicted survival probability at each time point. CRPS is always a number between 0 and 1, being 0 the best possible result. Survival prediction on the test cohort was performed using the *predict.rfsrc* function with default parameters. Harrel's concordance index (c-index) was used to assess model discriminative power on the bootstrapped training set and on the test set. C-index reflects to what extent a model predicts the order of events (e.g., deaths) in a cohort. C-indexes below 0.5 indicate poor prediction accuracy, c-indexes near 0.5 indicate random guessing and c-indexes of 1 represent perfect prediction.

Variable selection was selected by fitting age-adjusted cox regression models of overall survival in the training set. We initially analyzed different sets of genes in order to select the best input sets. Afterwards, we selected different sets of genes according to their multiple testing adjusted p-values; either False Discovery Rate (FDR) or Bonferroni-adjusted p-values below 0.01, 0.05, 0.1. Variable reduction was performed by iteratively removing those variables with low importance. Variable importance was calculated with the *vimp* function, and we iteratively removed those samples with negative or low weight (importance $< 1 \times 10^{-4}$). We replicated the models in the independent set.

RESULTS

Model Selection

We created six different survival models by taking into consideration different sets of transcripts at various thresholds of statistical significance. Importantly, all models achieved c-indexes > 0.67 in both the training and the validation sets. The best model contained 123 variables (absolute age, age ≥ 65, and the expression of 121 probes) (**Supplementary Table 1**). We name this model *Stellae-123* (ST-123). ST-123 achieved c-indexes of 0.7228 in the training set and 0.6988 in the test set, indicating a high reproducibility of the personalized risk prediction (**Figures 1A, B**). Apart from age, the most important variables in ST-123 were the expression of *KDM5B* and *LAPTM4B* (**Supplementary Table 2**).

Performance of ST-123 Over Time

The predictive accuracy of ST-123 was sustained over time, as reflected by the similar c-indexes obtained at 5 years of follow-up. Furthermore, mortality predictions were more accurate among those patients at high-risk of death according to the model. Additionally, CRPS plots indicate a rapid loss of predictability in the first days after diagnosis followed by an stabilization of survival predictions over time (**Figure 1C**). Indeed, c-indexes were less predictive for patients who died in the first days since diagnosis (**Table 1**). (ST-123 could clearly stratify patients in 4 different quartiles with different and reproducible mortality rates (**Figures 2A, B**).

ST-123 Performance in Patients With High-Risk Mutations

We tested the performance of the classifier in patients with high risk mutations in the TCGA cohort (validation set). We selected cases harboring *TP53* mutation and/or deletion (16 patients), *RUNX1* mutation (14 patients) and *ASXL1* mutation (6 patients) (1). We didn't include FLT3-ITD cases because this mutation type was not reported in the TCGA cohort. Using two-year survival predictions, we could observe a statistically significant association with overall survival within these high-risk patients (p-value 4.04 x 10^{-4}, age-adjusted p-value 2.16x 10^{-3}) (**Figure 2C**). Indeed, both cox models achieved high concordance measures: c-indexes of 0.774 (standard error 0.052) and 0.767 (standard error 0.042), respectively.

DISCUSSION

The concordance measure of ST-123 is substantially superior to that described for the ELN risk classification (0.59) (6), which suggests an increased performance. Furthermore, ST-123 seems particularly useful to better stratify high-risk patients. These results are in line with other ML-based models of survival based on mutational data described so far (8), which emphasizes the possibility of improving risk stratification in AML by implementing artificial intelligence. Although other prognostic models, such as that published by *Sherve et al.* (6), report a higher predictive accuracy, these also incorporate several baseline clinical variables. Indeed, CRPS plots indicate that the highest loss of prediction accuracy in our model occurs early after disease diagnosis, which reflects the fact that clinical variables are the main predictors of early mortality in AML (8). On the contrary, we can see how the CRPS curve of our model rapidly reaches a plateau, maintaining values ∿0.15 for long periods of time, reflecting the stability of the predictions over time. It seems reasonable that the incorporation of clinical variables (such as performance status and leukocytosis) will improve the overall predictive accuracy of ST-123 by identifying those patients at risk of early death. On the contrary, little contribution of mutations to gene expression patterns has been observed in myelodysplastic syndromes, and this issue needs to be addressed in AML (11).

It has been hypothesized that a better risk stratification based on the integration of different layers of clinical and biological complexity could, for example, reduce the number of allogeneic stem cell transplants performed in patients with AML by 20-25% while maintaining the same overall survival rate as that with the current treatment recommendations (8). In the same line, high-risk older patients might benefit from innovative drugs and drug combinations (12). This could help to restrict the most aggressive or innovative treatments to those AML patients who really are at the highest risk of death. Therefore, it becomes increasingly important to incorporate these new personalized models of survival in the risk stratification of AML, rather than keep relying on imperfect

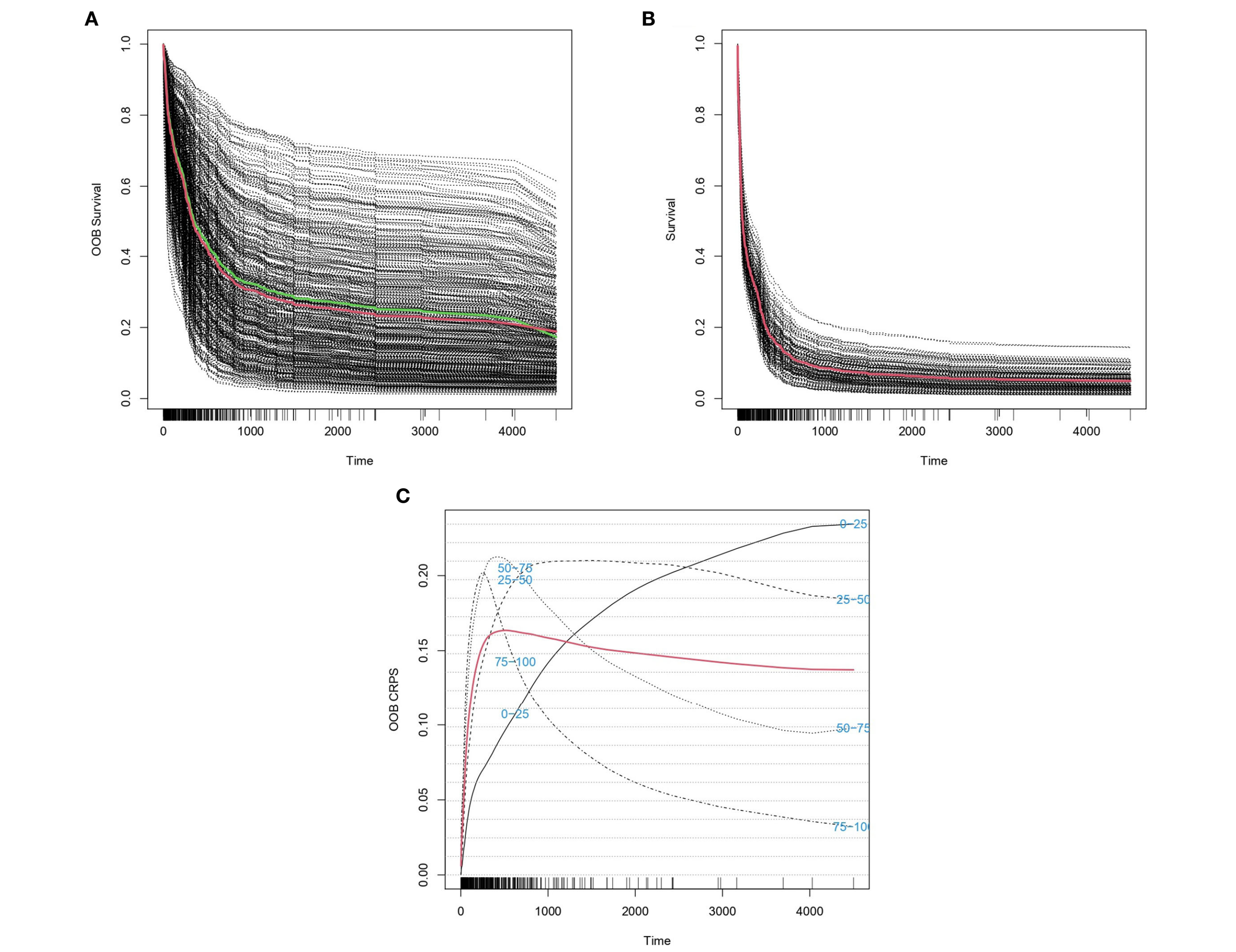

FIGURE 1 | Predicted individual survival curves according to the best random forests model. **(A)** Out-of-bag survival curves predicted for patients within the training cohort. The thick red line represents overall ensemble survival and the thick green line indicates the Nelson-Aalen estimator. **(B)** Individual survival curves predicted for patients within the test cohort. The thick red line represents overall ensemble survival. **(C)** Representation of out-of-bag CRPS over time. Red line is the overall CRPS. Additionally, stratified CRPS by quarters of predicted ensemble mortality are provided. Vertical lines above the x axis represent death events.

predefined risk groups. There is a need to analyze all these predictors in systematic and independent cohorts in order to provide further evidence about their effectiveness, and also to potentially integrate their individual predictive accuracies in more precise estimations of patient outcomes (13). Importantly, accumulated experience indicates that dichotomized molecular classifiers in AML can be reproduced in an unbiased and independent approach (14), but further efforts need to be made to compare the performance of personalized predictors such as ours.

From a biological point of view, the most important genes in ST-123 were *KDM5B* and *LAPTM4B*. Not surprisingly, both genes are deeply vinculated with carcinogenesis. *KDM5B* (Lysine Demethylase 5B) encodes a master epigenetic regulator of H3K4 methylation that regulates the expression of several oncogenes and tumor suppressors during carcinogenesis (15). *KDM5B* downregulates the oncogenic potential of leukemic stem cells by inducing H3K4-specific demethylation in murine and human MLL-rearranged AML cells, thereby promoting cell differentiation (16). Indeed, the KDM5B protein is the target of various inhibitors that are under study for the treatment of cancer (17). In the same line, the oncogene *LAPTM4B* (Lysosomal Protein Transmembrane 4 Beta) plays several roles in carcinogenesis, such as promoting tumor growth and metastasis, inhibiting apoptosis, initiating autophagy and driving multidrug resistance mechanisms (18). Interestingly, the expression of *LTPM4B* has been shown to be prognostic in myelodysplastic syndromes and in different types of solid tumors (19, 20).

In conclusion, our results indicate that survival of patients with AML can be predicted by applying ML tools to

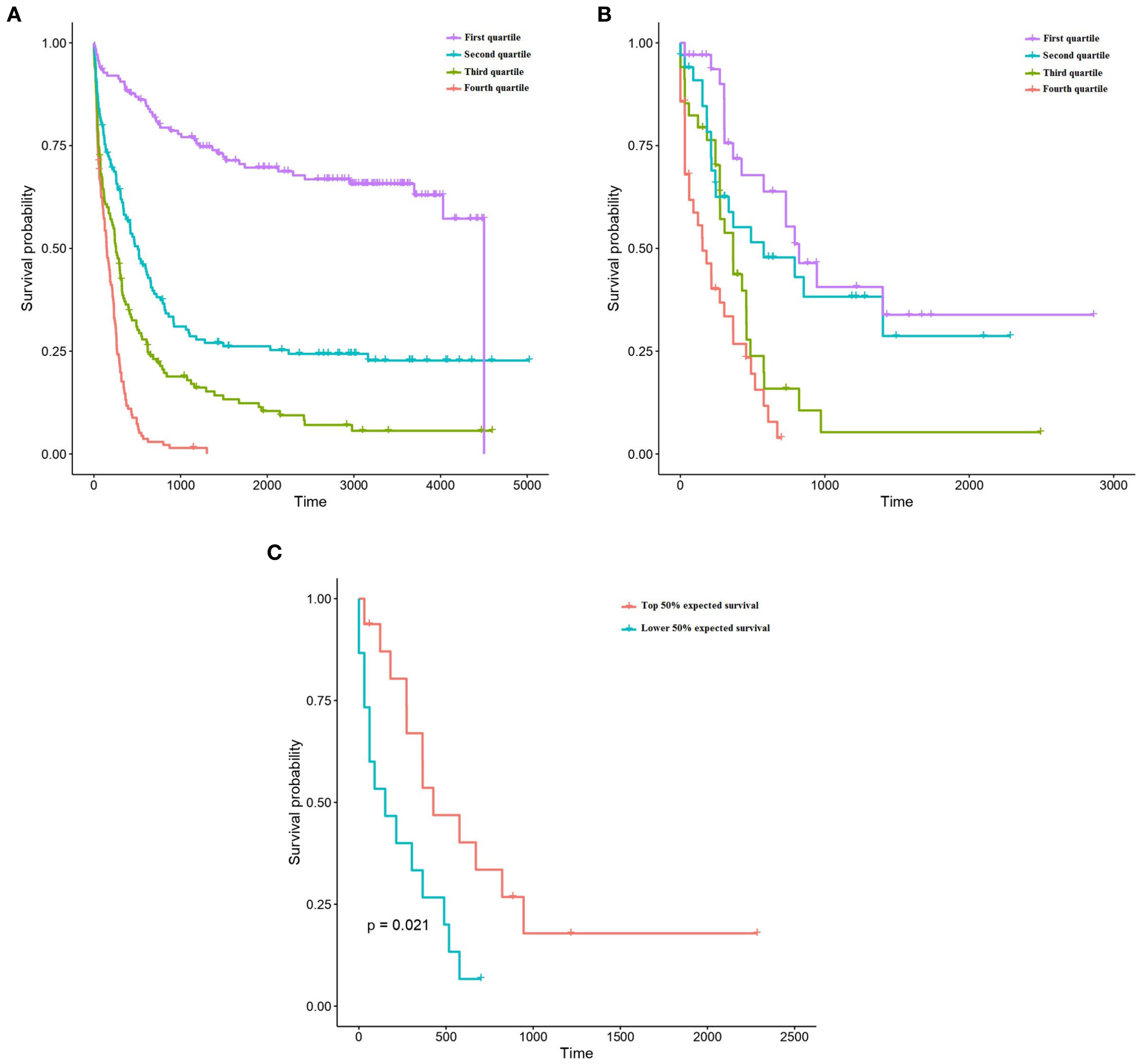

FIGURE 2 | Kaplan-Meyer plots representing the survival of patients depending on their classification to each quartile of predicted survival by ST-123 in the training **(A)** and validation **(B)** cohorts. **(C)** Kaplan-Meyer plots representing the outcomes of patients affected by high-risk mutations (*TP53* mutation/deletion, *ASXL1* mutation or *RUNX1* mutation) depending on their classification to the upper or lower median of predicted mortality by ST-123 in the validation set.

TABLE 1 | C-indexes of ST-123 after restricting the analysis to different time points since diagnosis.

Days since diagnosis	C-index (Training set)	C-index (Test set)
<10	68.65	59.66
<20	70.13	59.54
<30	71.88	60.19
<60	73.86	62.13
<100	74.06	63.11
>100	71.94	69.17

transcriptomic data, and that such predictions are particularly precise among patients with high-risk mutations. The possibility to enrich transcriptomic models such as ours with clinical and mutational data will lead to a more precise and holistic prediction of AML survival, paving the way for the development of new personalized treatment strategies.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

AM had the idea and performed the research. AM, AP, MC, JD, MG, and BA analyzed the data and wrote the paper. NA, LB, RF, AA, MA, and EF analyzed the manuscript and made suggestions. CC, GM, PM, MM, and JL gave final consent for publication. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We'd like to thank the *Supercomputing Center of Galicia* (CESGA) for their support. The content of this paper is part of the doctoral thesis of APR to obtain a PhD at the Department of Medicine, University of Santiago de Compostela.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.657191/full#supplementary-material

Supplementary Table 1 | Characteristics of the different random forest models created at the different significance thresholds. Error rates (1-concordance) are provided for predictions of survival at 1, 2, 3, 4, 5 and full follow-up periods.

Supplementary Table 2 | Variables included in the final model along with their relative importance. For gene expression probes, we also report the type of input used (either continuous gene expression counts or *Mclust*-based binary clusterization).

REFERENCES

1. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* (2017) 129 (4):424–47. doi: 10.1182/blood-2016-08-733196
2. Cafri G, Li L, Paxton EW, Fan JJ. Predicting risk for adverse health events using random forest. *J Appl Stat* (2018) 45(12):2279–94. doi: 10.1080/02664763.2017.1414166
3. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* (2019) 380(14):1347–58. doi: 10.1056/NEJMra1814259
4. Marcucci G, Haferlach T, Dohner H. Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *J Clin Oncol* (2011) 29:475–86. doi: 10.1200/JCO.2010.30.2554
5. Dohner H, Estey EH, Amadori S, Appelbaum FR, Buchner T, Burnett AK, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* (2010) 115:453–74. doi: 10.1182/blood-2009-07-235358
6. Sherve J, Meggendorfer M, Awada H. *A Personalized Prediction Model to Risk Stratify Patients with Acute Myeloid Leukemia (AML) Using Artificial Intelligence*. (2019). Available at: https://ashpublications.org/blood/article/134/Supplement_1/2091/427866/A-Personalized-Prediction-Model-to-Risk-Stratify.
7. Patkar N, Shaikh AF, Kakirde C, Nathany S, Ramesh H, Bhanshe P, et al. A novel machine-learning-derived genetic score correlates with measurable residual disease and is highly predictive of outcome in acute myeloid leukemia with mutated NPM1. *Blood Cancer J* (2019) 9:79. doi: 10.1038/s41408-019-0244-2
8. Gerstung M, Papaemmanuil E, Martincorena I, Bullinger L, Gaidzik VI, Paschka P, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat Genet* (2017) 49:332–40. doi: 10.1038/ng.3756
9. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J* (2016) 8 (1):205–33. doi: 10.32614/RJ-2016-021
10. Ishwaran H, Kogalur U, Blackstone E, Lauer M. Random survival forests. *Ann Appl Statist* (2008) 2(3):841–60. doi: 10.1214/08-AOAS169
11. Gerstung M, Pellagatti A, Malcovati L, Giagounidis A, Porta MG, Jädersten M, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun* (2015) 6:5901. doi: 10.1038/ncomms6901
12. DiNardo CD, Pratz K, Pullarkat V, Jonas BA, Arellano M, Becker PS, et al. Venetoclax combined with decitabine or azacitidine in treatment-naive, elderly patients with acute myeloid leukemia. *Blood* (2019) 133(1):7–17. doi: 10.1182/blood-2018-08-868752
13. Li Z, Herold T, He C. Identification of a 24-gene prognostic signature that improves the European LeukemiaNet risk classification of acute myeloid leukemia: an international collaborative study. *J Clin Oncol* (2013) 31 (9):1172–81. doi: 10.1200/JCO.2012.44.3184
14. Wang M, Lindberg J, Klevebring G. Validation of risk stratification models in acute myeloid leukemia using sequencing-based molecular profiling. *Leukemia* (2017) 31(10):2029–36. doi: 10.1038/leu.2017.48
15. Xhabija B, Kidder BL. KDM5B is a master regulator of the H3K4-methylome in stem cells, development and cancer. *Semin Cancer Biol* (2019) 57:79–85. doi: 10.1016/j.semcancer.2018.11.001
16. Wong SH, Goode DL, Iwasaki M, Wei MC, Kuo HP, Zhu L, et al. The H3K4-Methyl Epigenome Regulates Leukemia Stem Cell Oncogenic Potential. *Cancer Cell* (2015) 28(2):198–209. doi: 10.1016/j.ccell.2015.06.003
17. Fu YD, Huang MJ, Guo JW, You YZ, Liu HM, Huang LH, et al. Targeting histone demethylase KDM5B for cancer treatment. *Eur J Med Chem* (2020) 208:112760. doi: 10.1016/j.ejmech.2020.112760
18. Meng Y, Wang L, Chen D, Chang Y, Zhang M, Xu JJ, et al. LAPTM4B: an oncogene in various solid tumors and its functions. *Oncogene* (2016) 35 (50):6359–65. doi: 10.1038/onc.2016.189
19. Wang YH, Lin CC, Yao CY, Hsu CL, Hou HA, Tsai CH, et al. A 4-gene leukemic stem cell score can independently predict the prognosis of myelodysplastic syndrome patients. *Blood Adv* (2020) 4(4):644–54. doi: 10.1182/bloodadvances.2019001185
20. Dong X, Tamura K, Kobayashi D, Ando N, Sumita K, Maehara T. LAPTM4B-35 is a novel prognostic factor for glioblastoma. *J Neurooncol* (2017) 132 (2):295–303. doi: 10.1007/s11060-017-2369-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Mosquera Orgueira, Peleteiro Raíndo, Cid López, Díaz Arias, González Pérez, Antelo Rodríguez, Alonso Vence, Bao Pérez, Ferreiro Ferro, Albors Ferreiro, Abuín Blanco, Fontanes Trabazo, Cerchione, Martinnelli, Montesinos Fernández, Mateo Pérez Encinas and Luis Bello López. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.