REGULAR ARTICLE

blood advances

A geno-clinical decision model for the diagnosis of myelodysplastic syndromes

Nathan Radakovich,[1,2] Manja Meggendorfer,[3] Luca Malcovati,[4] C. Beau Hilton,[1,2] Mikkael A. Sekeres,[5] Jacob Shreve,[6] Yazan Rouphail,[7] Wencke Walter,[4] Stephan Hutter,[4] Anna Galli,[4] Sara Pozzi,[4] Chiara Elena,[4] Eric Padron,[8] Michael R. Savona,[9,10] Aaron T. Gerds,[1] Sudipto Mukherjee,[1] Yasunobu Nagata,[11] Rami S. Komrokji,[8] Babal K. Jha,[11] Claudia Haferlach,[4] Jaroslaw P. Maciejewski,[11] Torsten Haferlach,[3] and Aziz Nazha[1]

[1]Leukemia Program, Department of Hematology and Medical Oncology, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH; [2]Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH; [3]MLL Munich Leukemia Laboratory, Munich, Bavaria, Germany; [4]Department of Hematology Oncology, Fondazione IRCCS Policlinico San Matteo, University of Pavia, Pavia, Italy; [5]Division of Hematology, Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL; [6]Department of Internal Medicine, Cleveland Clinic, Cleveland, OH; [7]College of Arts and Sciences, The Ohio State University, Columbus, OH; [8]Department of Malignant Hematology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL; [9]Department of Medicine and [10]Department of Pediatrics, Program in Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN; and [11]Department of Translational Hematology and Oncology Research, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH

Key Points

- We developed a machine learning–based model to assist in the differential diagnosis of myeloid malignancies.
- Our work also describes genotype-phenotype correlations in different myeloid malignancies.

The differential diagnosis of myeloid malignancies is challenging and subject to interobserver variability. We used clinical and next-generation sequencing (NGS) data to develop a machine learning model for the diagnosis of myeloid malignancies independent of bone marrow biopsy data based on a 3-institution, international cohort of patients. The model achieves high performance, with model interpretations indicating that it relies on factors similar to those used by clinicians. In addition, we describe associations between NGS findings and clinically important phenotypes and introduce the use of machine learning algorithms to elucidate clinicogenomic relationships.

Introduction

Myelodysplastic syndromes (MDS) and other myeloid neoplasms are primarily diagnosed based on morphological changes in the bone marrow.[1] The diagnosis can be challenging, especially in patients with pancytopenia with minimal dysplasia, and is subject to interobserver variability, with up to 30% to 40% disagreement in diagnosis.[2] This difficulty is magnified for patients in whom the expected karyotypic or morphologic bone marrow changes are absent, or in the case of unsuccessful biopsies (eg, due to marked hypocellularity or extensive myelofibrosis).[3,4] In addition, there is increasing recognition of the relationship between MDS, myeloid neoplasms, and the states of idiopathic cytopenia of undetermined significance (ICUS) and clonal cytopenia of unknown significance (CCUS).[5] ICUS describes single or multiple cytopenias without a clear etiology or known clonal mutations, even after bone marrow biopsy evaluation. CCUS describes clonal mutations found with cytopenia(s) that do not meet criteria for a World Health Organization–defined hematologic neoplasm.[6]

Next-generation sequencing (NGS) has identified somatic and germline mutations that play a role in myeloid neoplasms' pathophysiology, progression, and response to therapy.[7,8] Such findings may refine presumptive diagnoses, especially in the absence of morphologic or karyotypic information. They are not, however, sufficiently specific to render a definitive diagnosis on their own, particularly in the case of ICUS, which is a diagnosis of exclusion.[5]

Machine learning (ML) is a family of computational algorithms that extract information by learning from relationships, patterns, and trends in data.[9] ML can produce powerful, reliable, and reproducible

Submitted 15 March 2021; accepted 28 July 2021; prepublished online on *Blood Advances* First Edition 30 September 2021; final version published online 29 October 2021. DOI 10.1182/bloodadvances.2021004755.

For data sharing, please contact the corresponding author at nazhaa@ccf.org.

The full-text version of this article contains a data supplement.

predictive models based on large and complex datasets.[10] Although ML models were traditionally considered "black boxes" that achieved performance at the expense of scrutability, modern methods allow the extraction of the most relevant features, that in turn allow models to be exploited and assure their reproducibility and accuracy when applied in clinical settings.[11]

In this study, we describe the use of an interpretable, ML model to differentiate MDS from other myeloid malignancies and provide patient-specific personalized interpretation of predictions using clinical and mutational data from peripheral blood without relying on information from a bone marrow biopsy.

Methods

Patients

Combined genomic and clinical data were collected from 652 patients treated at the Cleveland Clinic (United States), 1509 from the Munich Leukemia Laboratory (Germany), and 538 from the University of Pavia (Italy) between 2004 and 2018 (2697 patients in total). Clinical and laboratory data were retrospectively collected for all patients. To assure the accuracy and reproducibility of the model, other myeloid malignancies that resemble the presentation of MDS and can be in some cases difficult to distinguish from MDS, such as chronic myelomonocytic leukemia (CMML), MDS/myeloproliferative neoplasm (MPN), polycythemia vera (PV), essential thrombocythemia (ET), primary myelofibrosis (PMF), ICUS, and CCUS, were included. Inclusion criteria for patients included the following: availability of peripheral complete blood count and differential, bone marrow examination, and performance of NGS at the time of diagnosis. All patients' diagnoses were made by multidisciplinary teams experienced in the diagnosis of MDS and related disorders. Histopathologic examination of bone marrow specimens was performed by experienced hematopathologists who were not associated with the study or its outcomes in accordance with 2016 World Health Organization criteria.[1] The study was approved by each institution's Internal Review Board in accordance with the Declaration of Helsinki.

DNA sequencing analysis

Targeted deep sequencing was performed on 38 genes that are commonly reported in commercial genomic panels and have been demonstrated to be clinically relevant to MDS and other myeloid malignancies (supplemental Table 1). All genomic data were obtained via commercially available, clinically approved sequencing platforms. Detailed descriptions of sequencing procedures are available in the supplemental methods.

Model building and statistical analysis

Descriptive statistics were used to summarize the cohort. The χ^2 and Mann-Whitney tests were used to compare categorical and continuous data between disease subtypes, respectively. Cooccurrence and mutual exclusivity of mutations between genes were assessed using Fisher's exact test with the Benjamini-Hochberg correction used to account for multiple hypothesis testing. Similarly, a corrected Fisher's exact test was used to evaluate cooccurrence vs mutual exclusivity between disease subtypes and mutations, as well as between select clinical characteristics and mutations.

A gradient boosted machine (GBM) ML strategy, which generates predictions based on the input of multiple individual decision trees, was selected for model development for its ability to effectively handle nonlinear relationships in data and deliver state-of-the-art, interpretable performance.[12] To both train the model on a multi-institutional cohort and assess its external validity in a separate, independent cohort of patients, patients from the Cleveland Clinic and the University of Pavia were used as a train/test cohort; after model generation, model performance was then evaluated separately on the Munich Leukemia Laboratory cohort (graphical depiction in supplemental Figure 1). Models were developed using random 80% to 20% train-test splitting in the train/test cohort, with fivefold cross-validation and bootstrapping used to estimate confidence intervals (all confidence intervals are 95% unless otherwise specified). GBM algorithms were applied to predict myeloid malignancy diagnosis and bone marrow blast percentage. More details regarding model building are available in the supplemental appendix.

The feature importance package SHAP was used to identify genomic/clinical variables that impacted models' prediction, to visualize the impact of each variable on phenotype, and to generate individual predictions with explanations.[13] Prediction performance was evaluated according to area under the receiver operating characteristic curve (AUROC; this statistic measures the tradeoff between true positive and false positive predictions, with 0.5 representing a random guess and 1.0 representing a perfect predictor), confusion matrices, and inspection of individual predictions. Data analysis, model generation, and model interrogation were all performed using open-source Python packages (details in supplemental methods). Extensive descriptive data analyses are provided in the supplemental appendix.

Results

Patient characteristics

Of 2697 patients, 1630 (60.4%) had MDS, 399 (14.8%) had CMML, 142 (5.3%) had ICUS, 93 (3.4%) had CCUS, 129 (4.8%) had MDS/MPN, 41 (1.5%) had PV, 52 (1.9%) had ET, and 95 (3.5%) had primary PMF (Table 1). The clinical characteristics for the entire cohort and for each disease category are summarized in Table 1. Briefly, compared with MDS, the median age for MDS-MPN/CMML patients was older (70 vs 69 years; $P = .043$), equal for CCUS (68.4 years; $P = .948$), younger for ICUS (56 years; $P < .001$), and younger for MPNs (62; $P < .001$). Significant differences in the clinical and karyotypic data exist between MDS other cohorts (Table 1). (All P values are summarized in supplemental Table 1 in the supplemental appendix).

Molecular characteristics. To identify molecular signatures associated with each disease phenotype and to assure the reproducibility of our results across commercially available genomic panels, we focused all analyses on the 24 most common genes that were mutated in at least 2% of patients in our cohort (Figure 1A). These mutations are commonly included in all commercial laboratories and correlate with prior published reports.[14,15] We identified at least 1 mutation of these genes in 1711 patients (79%) with a median number of 2 mutations per sample (Q1: 1 mutation, Q3: 3 mutations).

The top mutated genes in MDS were *SF3B1* (26.5%), *TET2* (25.3%), and *ASXL1* (19.3%), differing in frequency from those in

Table 1. Cohort demographics, laboratory parameters, and cytogenetic variables

	MDS	MDS-MPN/CMML	MPN	ICUS	CCUS
			Mean (2.5th-97.5th percentile)		
Age, y	69.0 (41.00-85.70)	70.0 (43.00-85.86)	62.2 (28.35-83.90)	56.0 (22.44-84.55)	68.4 (40.55-85.04)
WBC, 10^9/L	5.8 (1.29-20.23)	19.8 (2.27-84.83)	15.5 (2.99-61.99)	4.4 (1.80-10.71)	4.7 (1.80-10.84)
Hemoglobin, g/dL	10.1 (6.80-14.10)	11.1 (7.00-15.39)	12.6 (7.19-20.70)	12.2 (7.10-16.09)	11.9 (7.57-15.01)
Platelets, 10^{12}/L	183.1 (15.00-650.15)	182.6 (15.00-735.75)	362.5 (10.13-1069.30)	159.8 (19.52-387.32)	142.5 (18.10-350.50)
ANC, 10^9/L	3.1 (0.27-12.07)	9.4 (0.59-37.21)	10.3 (1.34-42.70)	2.5 (0.19-7.34)	2.5 (0.52-7.45)
ALC, 10^9/L	1.0 (0.04-3.63)	2.6 (0.44-7.85)	2.3 (0.31-6.54)	1.5 (0.46-3.25)	1.6 (0.45-3.92)
AMC, 10^9/L	0.3 (0.00-1.50)	4.8 (0.08-26.28)	0.9 (0.00-4.40)	0.4 (0.02-1.08)	0.5 (0.07-1.38)
BM blast, %	5.0 (0.00-17.00)	5.5 (0.00-19.00)	1.5 (0.00-6.70)	1.6 (0.00-4.50)	1.8 (0.00-4.60)
Peripheral blasts, 10^9/L	0.3 (0.00-3.00)	1.6 (0.00-12.00)	1.4 (0.00-8.32)	0.0 (0.00-0.00)	0.0 (0.00-0.00)
			Number (%)		
Female	1005 (37.26)	392 (14.53)	89 (3.30)	67 (2.48)	57 (2.11)
Normal karyotype	810 (30.03)	103 (3.82)	104 (3.86)	44 (1.63)	30 (1.11)
Chr 5 abnormality	135 (5.01)	6 (0.22)	4 (0.15)	0 (0.00)	0 (0.00)
Chr 7 abnormality	73 (2.71)	18 (0.67)	5 (0.19)	0 (0.00)	0 (0.00)
Complex karyotype	105 (3.89)	12 (0.44)	5 (0.19)	0 (0.00)	1 (0.04)
			P		
Age, y	reference	0.043	5.04E-12	1.27E-17	0.948
WBC, 10^9/L	reference	5.47E-122	3.94E-49	0.13	0.579
Hemoglobin, g/dL	reference	7.32E-21	7.99E-19	3.13E-25	2.86E-17
Platelets, 10^{12}/L	reference	0.245	4.60E-19	0.727	5.13E-01
ANC, 10^9/L	reference	2.28E-79	2.23E-54	0.552	5.06E-01
ALC, 10^9/L	reference	6.23E-115	2.78E-34	4.27E-19	7.75E-15
AMC, 10^9/L	reference	1.77E-194	3.03E-34	8.28E-15	1.58E-13
BM blast %	reference	0.659	2.17E-24	1.43E-07	2.13E-05
Peripheral blasts, 10^9/L	reference	6.35E-54	2.58E-43	0.08	1.09E-01
Female	reference	0.468	2.94E-08	0.000987	9.69E-01
Normal karyotype	reference	1.10E-45	0.819	2.77E-05	2.00E-03
Chr 5 abnormality	reference	3.26E-10	0.001	0.000666	7.00E-03
Chr 7 abnormality	reference	0.118	0.199	0.019	6.90E-02
Complex karyotype	reference	2.89E-05	0.026	0.003	6.10E-02

ALC, absolute lymphocyte count; AMC, absolute monocyte count; ANC, absolute neutrophil count; BM, bone marrow; Chr, chromosome; WBC , white blood cell.

MDS-MPN/CMML (*TET2* 46.3%, *ASXL1* 33.7%, *SRSF2* 28.8%), CCUS (*TET2* 40.9%, *DNMT3A* 26.9%, *ASXL1* 19.4%), and MPN (*JAK2* 63.6%, *ASXL1* 26.6%, *TET2* 14.0%) (Figure 1B). As expected, *JAK2* mutations were the most common mutations in MPNs (Figure 1B). Gene mutation frequencies correlated with what has been reported in previous studies.[14-16]

Correlation of molecular mutations among each other and with clinical variables in each disease phenotype. We evaluated the correlation between the mutations with each other, disease phenotype, and clinical variables (Figures 1C and 2). As expected, strong correlations were observed between *SF3B1* and MDS-RS, *TET2/SRSF2* and CMML, and *JAK2* mutations and MPNs (Figure 2). *TP53* and *RUNX1* mutations were associated with MDS with excess blasts, whereas *SF3B1* was mutually exclusive with this disease phenotype; the converse was observed for MDS with ring sideroblasts, which was also associated with a favorable outcome.

The correlation among mutations differed in each phenotype (Figure 2). Briefly, in MDS, *ASXL1* mutations were associated with mutations in other genes involved in epigenetic modification, such as *SRSF2*, *CBL*, and *STAG2*. Mutual exclusivity was seen between *SF3B1* and *SRSF2*, *RUNX1*, *U2AF1*, *STAG2*, *TP53*, *NRAS*, *IDH2*, and *NPM1*. Mutual exclusivity was also observed between *TP53* and *ASXL1*, *SF3B1*, and *SRSF2*. Among patients with CMML/CCUS/MDS-MPN, cooccurrence was observed between *TET2* and *SRSF2*, between *ASXL1* and *RUNX1*, and between *ASXL1* and *EZH2* (Figure 2: no significant co-occurrence or mutual exclusivity was observed in mutations in the MPN category, ET/PMF/PV).

Interesting correlations were also observed between mutations and the clinical variables (Figure 2). As expected, *TP53* correlated strongly with complex karyotype, chromosome 5 abnormalities, and chromosome 7 abnormalities (Figure 2). Conversely, *SF3B1* mutations were associated with normal karyotype and bone marrow blasts <10% while demonstrating mutual exclusivity with complex

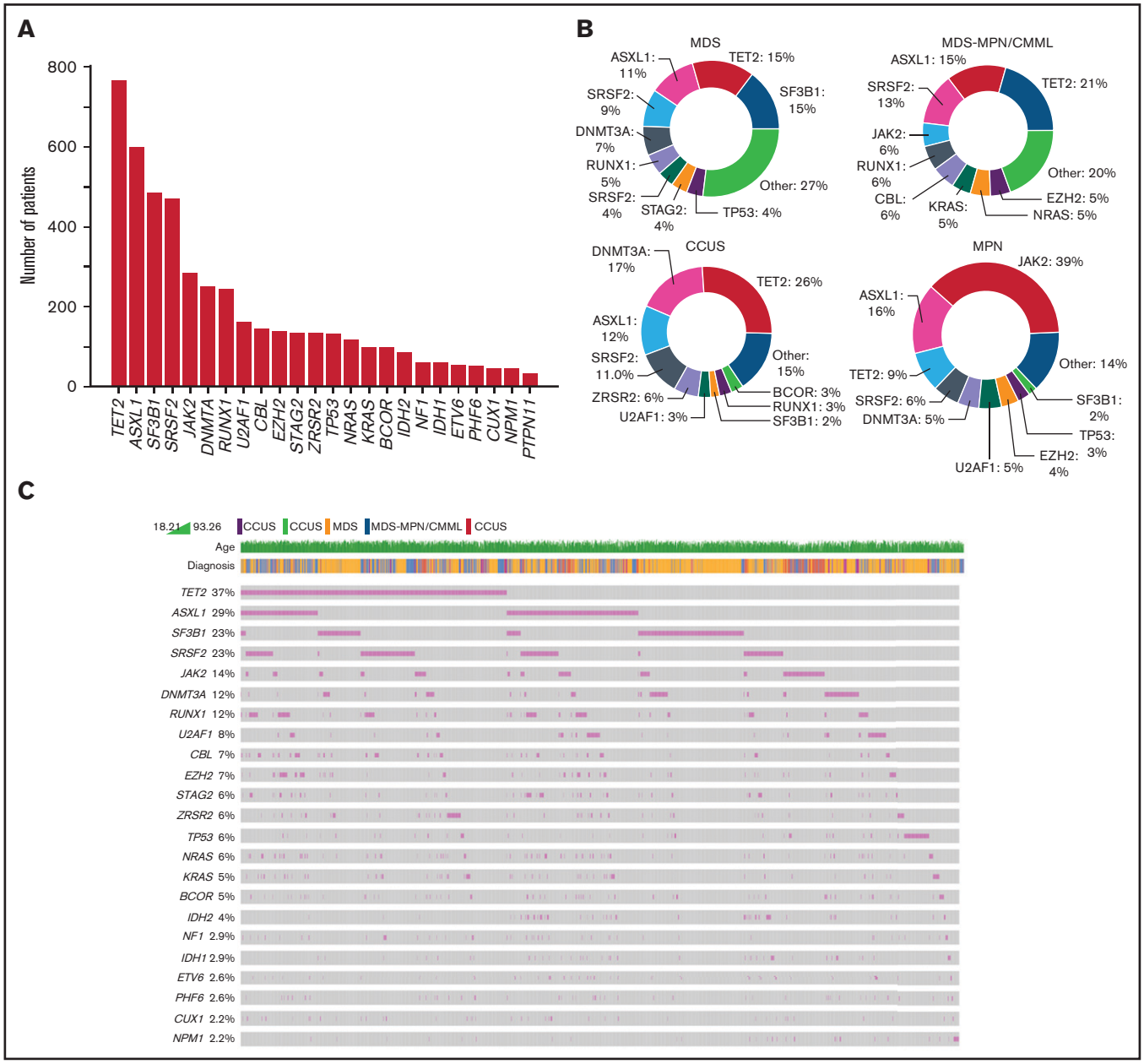

Figure 1. Cohort genomic characteristics. (A) Top mutated genes in cohort. (B) Mutation frequency by disease subtype. (C) Cohort-wide oncoprint.

karyotype, chromosome 7 abnormalities, and low platelets. These observations reflect the favorable prognosis associated with *SF3B1* mutations.[17]

Model development

Fifteen genomic/clinical variables were included in the final model that was used to accurately diagnose MDS, in distinguishing it from other conditions (supplemental Table 2). The model explanation algorithm SHAP was used to identify the variables with the most substantial impact on model predictions; to simplify model use and reduce the degree of overfitting, the top 15 variables were retained for the final model. Overall, the most important variables were (in descending order of importance) as follows: number of mutations (greater mutation number being associated with a prediction of MDS and lower numbers more associated with predictions of ICUS or CCUS), percentage of blasts in peripheral blood (most associated with MDS/MPN-CMML), absolute monocyte count (most associated with MDS/MPN-CMML and a negative predictor for MDS), *JAK2* status (most predictive for MPN), hemoglobin (elevated or normal hemoglobin associated with MPN, and relatively mild anemia associated with CCUS/ICUS), absolute basophil count, age, absolute eosinophil count, absolute lymphocyte count, absolute neutrophil count, *KRAS* status, *SF3B1* status, and platelet count (Figure 3A). The most important variables for each class, along with the relative impact of high or low values for those variables, are shown in

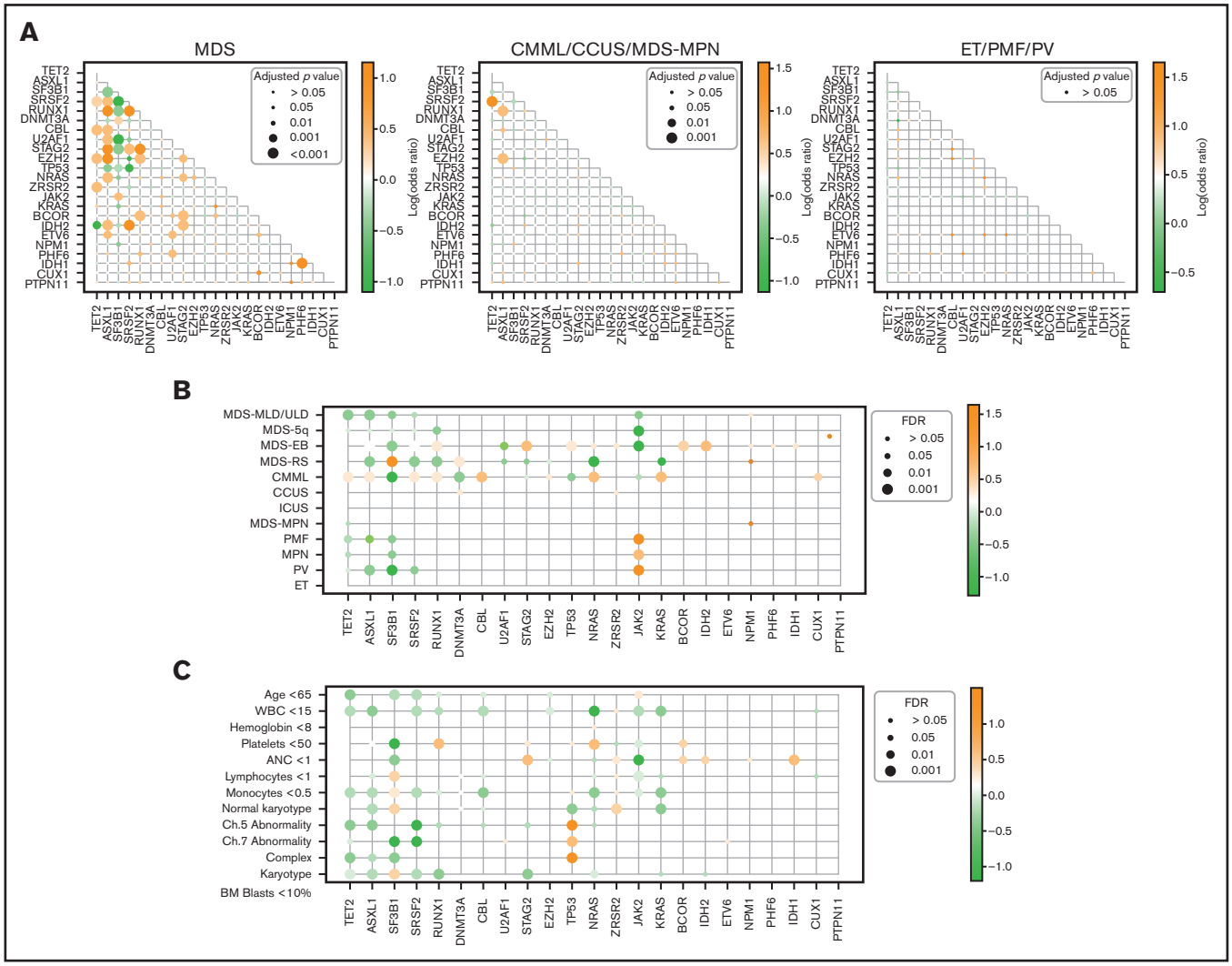

Figure 2. Cohort geno-geno and geno-clinical correlations. (A) Coexpression vs exclusivity between mutations by disease subtype. (B) Disease-phenotype correlations. (C) Mutation-phenotype correlations.

individual summary plots (representative examples shown in Figure 3B; for instance, in 1 patient correctly identified as having MDS/MPN-CMML, the model's prediction is driven by monocytosis for the example patient, but the patient's thrombocytosis, leukocytosis, and elevated absolute neutrophil count [ANC] also factor into the prediction).

Model diagnostic performance

When applied to the test and validation cohorts, model performance was as follows (a perfect predictor has an AUROC of 1, and AUROC > 0.90 are generally considered excellent): AUROC of 0.951 (0.934 to 0.966) for test cohorts, and AUROC of 0.926 (0.916 to 0.937) for the training cohorts. Inspection of cohort-level feature importance revealed which patient characteristics were most influential for the diagnosis of each disease subtype (Figure 3A). In addition to cohort-level model explanations, quantitative differential diagnoses were produced for each patient in the validation cohort (examples shown in Figure 3B).

Other predictions

A GBM model was also used to predict important clinical and genetic characteristics based on one another (eg, using clinical and karyotypic data to predict individual mutation status in the absence of NGS data, and vice versa). When NGS data and patient sex were used as inputs, complex karyotype was predicted with an AUROC of 0.821, normal karyotype with an AUROC of 0.790, abnormal karyotype with an AUROC of 0.761, age with an AUROC of 0.675, and severe pancytopenia (defined as hemoglobin <8 g/dL, platelets <50 000/uL, and ANC <1000/uL) with an AUROC of 0.700. Cohort-level explanations for model predictions reflected known associations in myeloid malignancies, such as concordance between *TP53* mutations and complex karyotype, *TET2* mutations and age >65 years, and exclusivity between *SF3B1* mutations and either severe pancytopenia and complex karyotype. Using karyotype and peripheral blood count data, the model was able to predict *TP53*, *SF3B1*, and *IDH* status with an AUROC of 0.87, 0.78, and 0.90, respectively. Feature importance graphs as calculated by Shapley values are provided in Figure 4.

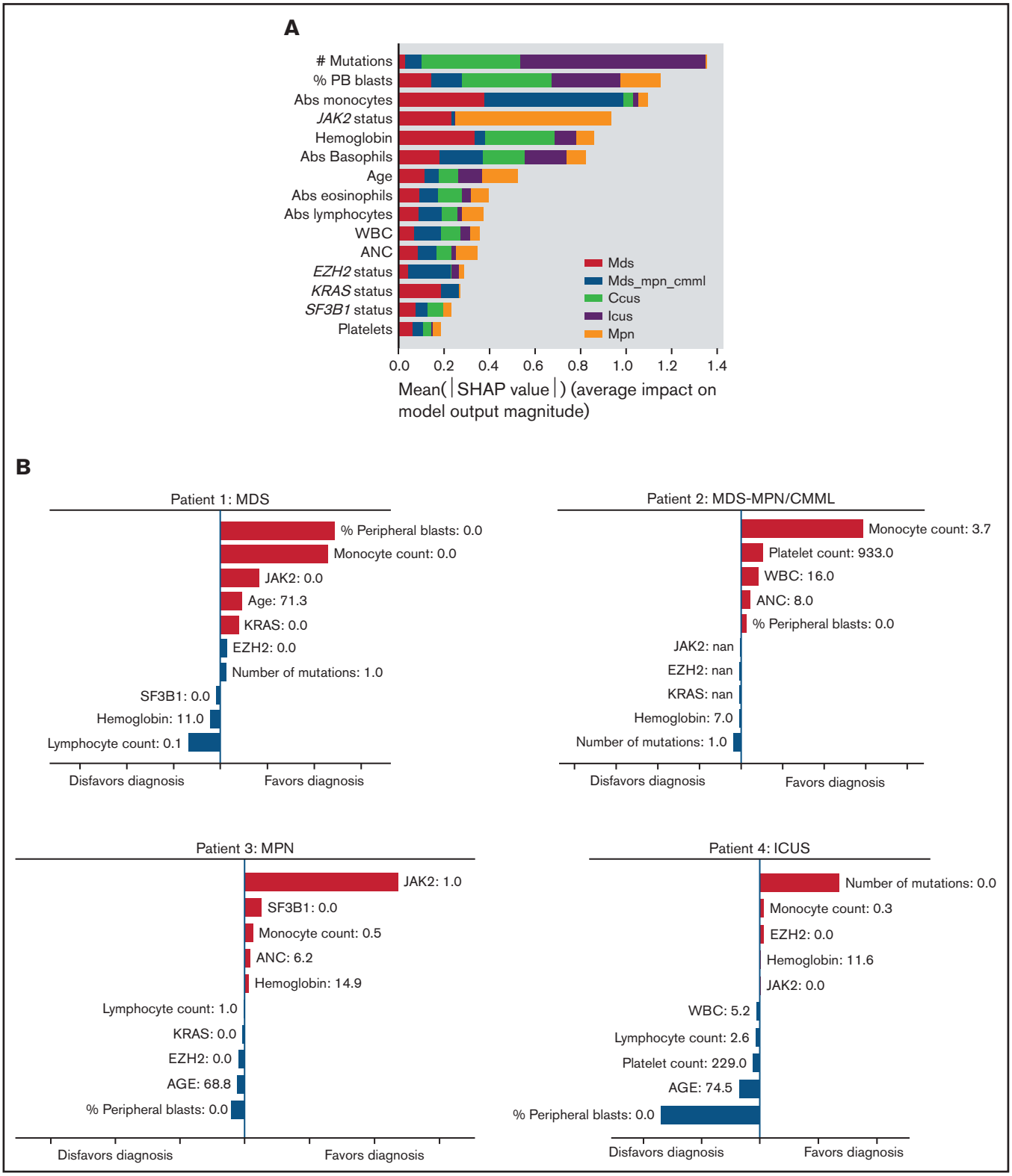

Figure 3. Model feature importance. (A) Global model feature importance. (B) For individuals, features in blue decrease the probability of a given diagnosis, while features in red increase it. The size of the red or blue bar for a given feature reflects its relative contribution. The probability for each diagnosis is given in the "output value," as a number between zero and one.

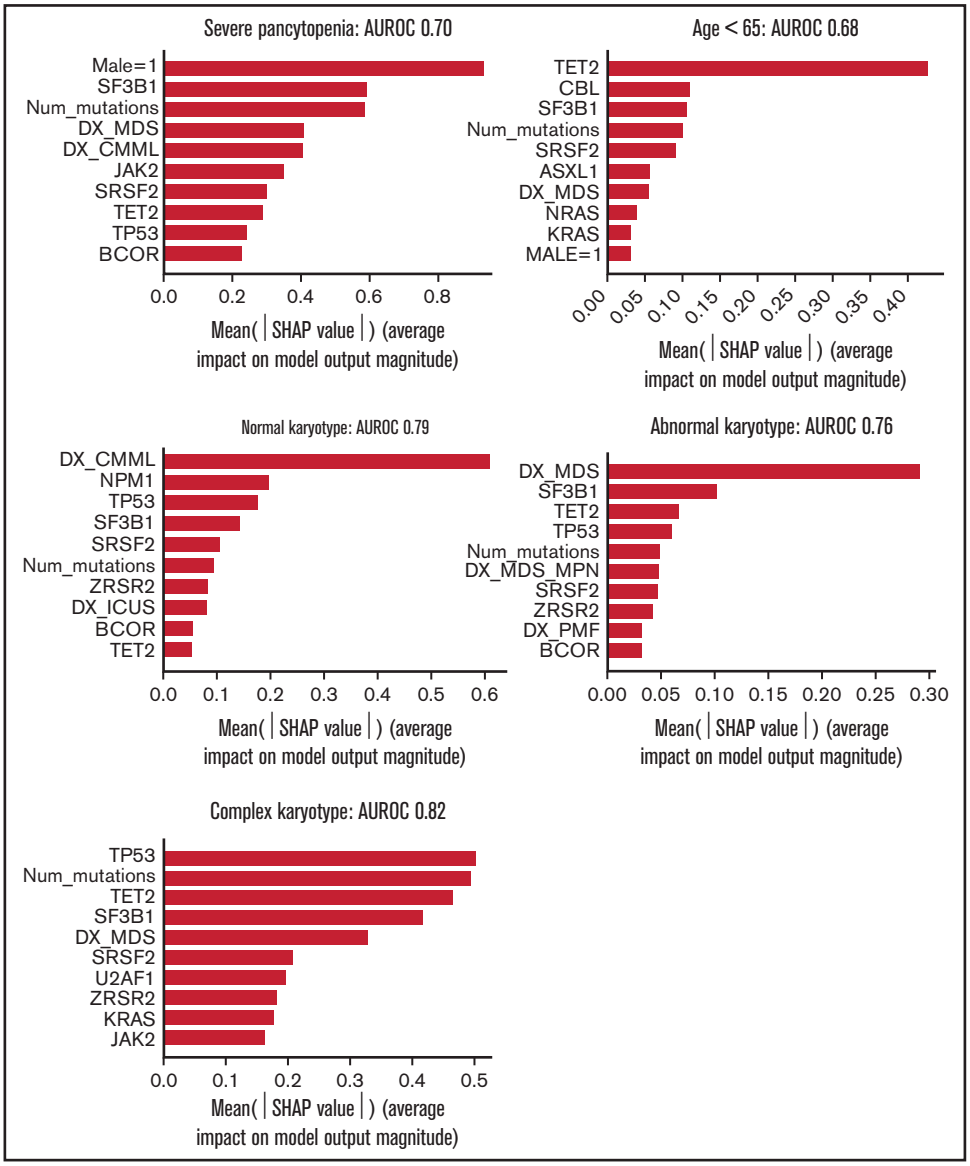

Figure 4. Feature importance for prediction of clinical variables based on NGS data. Importance plots are shown for severe pancytopenia, age <65 years, normal karyotype, abnormal karyotype, and complex karyotype. Bars for each feature represent relative importance of given feature for predicting a clinical characteristic.

Discussion

The differential diagnosis of MDS includes several similar-appearing neoplasms that can be difficult to distinguish from one another, even for physicians who specialize in such disorders. We describe an ML-based approach to aid in the differential diagnosis of myeloid malignancies using data collected exclusively from peripheral blood in patients receiving treatment at tertiary malignant hematology practices. The model developed with this approach was able to generate highly accurate predictions in the absence of any data collected from patients' bone marrow.

As described by Shapley values, the most pivotal features for differential diagnosis were primarily clinical, with the exception of the total number of mutations, as well as *JAK2*, *KRAS*, and *SF3B1* status. This is unsurprising, as few of the mutations studied are considered exclusive hallmarks of a particular malignancy and thus would not be expected to be particularly informative for differential diagnosis. The predictive genomic features identified in our model are exceptions, as they are documented hallmarks of particular conditions (ICUS, MPNs, and MDS, for absence of any mutations, *JAK2* status, and *SF3B1* status, respectively). Beyond aggregate feature importance, the model was able to generate sensible explanations for individual differential diagnoses. This is critical, as clinically applicable artificial intelligence must be interpretable in order for it to interface with the other lines of evidence and data sources that clinicians use to deliver patient care.

A diagnostic model such as this one has several potential applications for clinicians. Although histopathological evaluation remains a linchpin in the diagnosis of myeloid malignancies, poor sample

quality, fibrosis, or hypocellularity often serve as barriers to reliable evaluation.[16] Models capable of making predictions exclusive of histopathological information may offer clarity in such situations, or when tissue specimens are adequate but ambiguous, and patients are unwilling to undergo repeat bone marrow assessments. In addition, although this model was developed based on patients from tertiary care centers (and thus likely represents an older and more advanced state of disease compared with myeloid malignancies in aggregate), it may represent a useful approach for general oncologists, either as an additional data point to consider when diagnosing a condition or when contemplating referrals to specialists in hematologic malignancies. Previous work such as that by Moraes et al, who apply ML to the differential diagnosis of lymphoproliferative disorders via flow cytometry,[18] highlights the utility of these techniques for making great use of existing data and providing complementary means of rendering diagnoses. The addition of tools such as this to electronic health records could aid diagnostic decision making by suggesting likely diagnoses as well as providing rationales for competing differential diagnoses.

We also describe, using only patient age and NGS data, the prediction of several clinically relevant patient characteristics, such as abnormal or complex karyotypes and severe pancytopenia. Although some of these characteristics (such as complete blood counts) are not intrinsically difficult to obtain, these findings are hypothesis generating in determining how different genetic alterations contribute to the varied presentations of MDS and related malignancies. In addition, the ability to predict certain disease phenotypes, such as cytopenias leading to frequent transfusions or high infection risk, may inform expectations about the degree of supportive care required for patients, or even the timing and relative risk of stem cell transplant.

Our study also investigated the ability to predict clinically meaningful single-gene alterations such as *TP53* and *SF3B1* status. In addition to its hypothesis-generating role, being able to predict such key genomic alterations based off readily available data may play a role in upfront decision making in patients with myeloid malignancies, as the presence of certain mutations such as *TP53*, *IDH1*, and *SF3B1* has established prognostic and therapeutic ramifications.[17,19,20] This could affect the decision to pursue targeted therapies, or in the case of *TP53* mutations, eschew chemotherapy. Such predictions would eventually have to be confirmed by NGS, in clinical practice obtaining NGS data may take weeks, delaying treatment decisions. Importantly, although our approach to predicting phenotype-genotype correlations is new, it successfully identifies known relationships in myeloid malignancies, such as the well-established correlation between high numbers of chromosomal abnormalities and mutant *TP53* status.[21,22] Although observations from this approach should be considered hypothesis generating, they offer an intriguing means of exploring genotype-phenotype relationships.

Our study has some limitations. Although large by the standard of studies investigating malignancies, the size of our dataset precluded rarer mutations from being incorporated into the model. Thus, although our work does positively identify several predictive factors, it cannot exclude the significance of other, infrequent mutations. It is also important to note that the diagnoses associated with our training data reflect contemporaneous diagnostic standards and definitions of each disease subtype. As such, the estimates of model performance described assume it will be applied in settings where similar diagnostic criteria are used. Furthermore, prospective validation will provide important confirmation of the diagnostic accuracy of our model, as well as confirmation of its ability to assist with ambiguous or difficult diagnoses. Finally, although our model was able to accurately make use of genomic data obtained with different platforms and protocols, it is unclear if a model using more granular data such as variant allele frequency would be similarly robust across platforms. Although beyond the scope of this work, it will remain vital that diagnostic practices between institutions are continually examined, especially as new diagnostic modalities emerge.

In summary, we describe an ML-based approach to the diagnosis of myeloid malignancies absent the data typically obtained from a bone marrow biopsy. Our model's findings and predictions are consistent with known hallmarks of these diseases and demonstrate the potential utility of ML-based approaches as an additional tool in the upfront evaluation of these diseases.

Authorship

Contribution: N.R. generated ML models, performed primary analysis, and wrote the original manuscript; C.H. generated models, performed analysis, and wrote the original manuscript; A.N. conceptualized the project, assisted in data interpretation, provided data, and edited the manuscript; M.M., L.M., M.A.S., W.W., S.H., A.G., S.P., C.E., E.P., M.R.S., A.T.G., S.M., Y.N., R.S.K., B.K.J., C.H., J.P.M., and T.H. all contributed both primary data and clinical data as well as the bioinformatics pipeline used to analyze NGS data; and J.S. and Y.R. both assisted with code and helped edit the manuscript.

Conflict-of-interest disclosure: S.M. declares the following disclosures: advisory board participation for Celgene/Acceleron, Bristol Myers Squib, Novartis, Blueprint Medicines, Genentech, and AbbVie; receipt of honoraria from Aplastic Anemia and MDS International Foundation, Celgene (now Bristol Myers Squib), Bristol Myers Squib, McGraw Hill Hematology Oncology Board Review, Partnership for Health Analytic Research, LLC (PHAR, LLC); receipt of consultancy fees from BioPharm, Celgene, Novartis, BMS; and receipt of research funding from BMS (formerly Celgene), Novartis, and Jazz Pharmaceuticals. The remaining authors declare no competing financial interests.

ORCID profiles: L.M., 0000-0002-1460-1611; C.H., 0000-0002-1363-7452; Y.R., 0000-0002-1859-3225; W.W., 0000-0002-5083-9838; A.G., 0000-0002-4039-3756; M.R.S., 0000-0003-3763-5504; A.T.G., 0000-0002-3422-1309; R.S.K., 0000-0002-1876-5269; B.K.J., 0000-0002-7660-5255.

Correspondence: Aziz Nazha, Cleveland Clinic Center for Clinical Artificial Intelligence, Enterprise Analytics, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Department of Hematology and Medical Oncology, Taussig Cancer Institute, Cleveland Clinic, Desk R35 9500 Euclid Ave, Cleveland, OH 44195; e-mail: nazhaa@ccf.org.

References

1. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood.* 2016;127(20):2391-2405.
2. Font P, Loscertales J, Benavente C, et al. Inter-observer variance with the diagnosis of myelodysplastic syndromes (MDS) following the 2008 WHO classification. *Ann Hematol.* 2013;92(1):19-24.
3. Sasada K, Yamamoto N, Masuda H, et al; Kyushu regional department of the Japanese Society of Laboratory Hematology. Inter-observer variance and the need for standardization in the morphological classification of myelodysplastic syndrome. *Leuk Res.* 2018;69:54-59.
4. Shaver AC, Seegmiller AC. Nuances of morphology in myelodysplastic diseases in the age of molecular diagnostics. *Curr Hematol Malig Rep.* 2017;12(5):448-454.
5. Steensma DP, Bejar R, Jaiswal S, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood.* 2015;126(1):9-16.
6. Valent P, Orazi A, Steensma DP, et al. Proposed minimal diagnostic criteria for myelodysplastic syndromes (MDS) and potential pre-MDS conditions. *Oncotarget.* 2017;8(43):73483-73500.
7. Nazha A, Bejar R. Molecular data and the IPSS-R: how mutational burden can affect prognostication in MDS. *Curr Hematol Malig Rep.* 2017; 12(5):461-467.
8. Haferlach T. Molecular genetics in myelodysplastic syndromes. *Leuk Res.* 2012;36(12):1459-1462.
9. Meyer R, Obermayer K. pypet: a python toolkit for data management of parameter explorations. *Front Neuroinform.* 2016;10:38.
10. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J.* 2014;13:8-17.
11. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng.* 2018;2(10):749-760.
12. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30. Curran Associates, Inc. 3146-3154.
13. Lundberg S, Lee S. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30. Curran Associates, Inc. 4765-4774.
14. Larsson CA, Cote G, Quintás-Cardama A. The changing mutational landscape of acute myeloid leukemia and myelodysplastic syndrome. *Mol Cancer Res.* 2013;11(8):815-827.
15. Pellagatti A, Boultwood J. The molecular pathogenesis of the myelodysplastic syndromes. *Eur J Haematol.* 2015;95(1):3-15.
16. Senent L, Arenillas L, Luño E, Ruiz JC, Sanz G, Florensa L. Reproducibility of the World Health Organization 2008 criteria for myelodysplastic syndromes. *Haematologica.* 2013;98(4):568-575.
17. Malcovati L, Stevenson K, Papaemmanuil E, et al. SF3B1-mutant MDS as a distinct disease subtype: a proposal from the International Working Group for the Prognosis of MDS. *Blood.* 2020;136(2):157-170.
18. Moraes LO, Pedreira CE, Barrena S, Lopez A, Orfao A. A decision-tree approach for the differential diagnosis of chronic lymphoid leukemias and peripheral B-cell lymphomas. *Comput Methods Programs Biomed.* 2019;178:85-90.
19. Montalban-Bravo G, Garcia-Manero G, Jabbour E. Therapeutic choices after hypomethylating agent resistance for myelodysplastic syndromes. *Curr Opin Hematol.* 2018;25(2):146-153.
20. Cai L, Zhao X, Ai L, Wang H. Role of TP53 mutations in predicting the clinical efficacy of hypomethylating therapy in patients with myelodysplastic syndrome and related neoplasms: a systematic review and meta-analysis. *Clin Exp Med.* 2020;20(3):361-371.
21. Rücker FG, Schlenk RF, Bullinger L, et al. TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome. *Blood.* 2012;119(9):2114-2121.
22. Becker H, Pfeifer D, Ihorst G, et al. Monosomal karyotype and chromosome 17p loss or TP53 mutations in decitabine-treated patients with acute myeloid leukemia. *Ann Hematol.* 2020;99(7):1551-1560.