Supplemental Materials

Differential diagnosis of bone marrow failure syndromes guided by machine learning

Fernanda Gutierrez-Rodrigues,[1,*] Eric Munger,[2,*] Xiaoyang Ma,[1] Emma M. Groarke,[1] Youbao Tang,[6] Bhavisha A. Patel,[1] Luiz Fernando B. Catto,[3] Diego V. Clé,[3] Marena R. Niewisch,[7] Raquel M. Alves-Paiva,[4] Flávia S. Donaires,[3] André Luiz Pinto,[3] Gustavo Borges,[3] Barbara A. Santana,[3] Lisa J. McReynolds,[7] Neelam Giri,[7] Burak Altintas,[7] Xing Fan,[11] Ruba Shalhoub,[1] Christopher M Siwy,[13] Carrie Diamond,[1] Diego Quinones Raffo,[1] Kathleen Craft,[8] Sachiko Kajigaya,[1] Ronald M. Summers,[6] Paul Liu,[8] Lea Cunningham,[8] Dennis D. Hickstein,[9] Cynthia E. Dunbar,[11] Ricardo Pasquini,[5] Michel Michels De Oliveira,[5] Elvira D. R. P. Velloso,[4,10] Blanche P. Alter,[7] Sharon A. Savage,[7] Carmem Bonfim,[5] Colin O. Wu,[12] Rodrigo T. Calado,[3] and Neal S. Young [1]

[1]Hematology Branch, National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), Bethesda, Maryland, USA. [2]Department of Bioinformatics and Computational Biology, George Mason University, Fairfax, Virginia, USA. [3]Department of Medical Imaging, Hematology, and Oncology, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil.[4] Israelita Albert Einstein, São Paulo, SP, Brazil.[5] Bone Marrow Transplantation unit, Federal University of Parana, Curitiba, Brazil. [6]NIH, Clinical Center, Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Bethesda, Maryland, United States.[7]Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI). [8]Translational and Functional Genomics Branch, National Human Genome Research Institute (NHGRI), Maryland, USA. [9]Experimental Transplantation and Immunology Branch, NCI, Maryland, USA. [10]Service of Hematology, Transfusion and Cell Therapy and Laboratory of Medical Investigation in Pathogenesis and Directed Therapy in Onco-Immuno-Hematology (LIM-31) HCFMUSP, University of Sao Paulo Medical School, Sao Paulo, Brazil. [11]Translational Stem Cell Biology Branch, NHLBI, NIH, Bethesda, MD, USA. [12]Office of Biostatistics Research, NHLBI, NIH, Bethesda, Maryland, USA.[13]NIH Clinical Center, Department of Clinical Research Informatics.

Contents

1. Quick guide in machine learning: concepts and applications

Machine-learning approaches are computer algorithms that learn from examples rather than a pre-established set of statistical rules, being a powerful tool to discover patterns within complex data. Therefore, machine learning has been applied in many studies for prediction, risk stratification, and image processing. An extensive review of machine learning can be found elsewhere.[1,2] Here, we aim to provide an overview of machine learning focused on some basic concepts required for the interpretation of our results.

1.1. Process overview

Most implementations of machine-learning methods share some common characteristics:

1) **Data processing:** machine-learning algorithms learn more effectively from large numbers of well-curated examples. These ***Examples*** are annotated according to their ***labels*** (an annotation that flags what is the target for prediction or pattern recognition) and contain a list of ***variables*** that will be used as input for the algorithm. Machine-learning approaches are "data-hungry", which means that it requires a large number of examples for accurate performance. In heterogeneous datasets, ideally, it is best if the number of examples with different labels is ***balanced.*** This improves the learned accuracy but may not necessarily represent the true disease prevalence in a population. Balancing the number of class label examples reduces the confounding impact of differences between datasets in the analysis. Inferred data or under-sampling are two common solutions for label skew.

2) **Development of a machine-learning algorithm. Training phase.** An algorithm will process a dataset and store the "rule learned from the data" in a ***model.*** In this work, two main types of machine-learning algorithms are used for data modeling: ***unsupervised and supervised learning***. Unsupervised learning focuses on pattern recognition and is used to cluster individual observations into bins with other observations with similar patterns without explicitly using the target label. Supervised learning is used for selecting the minimum number of variables to be used in an optimized classification model (feature selection or ***top predictors*** for a specific target). Supervised learning is also used to train and test a final classification model that can be used to label new observations.

3) **Development of a machine-learning algorithm. Validation phase.** A stored model will be used to predict labels in a dataset of new observations containing the same variables from the dataset used for training a model. If the rule learned by the model is successful, then most of the cases in the validation cohort will be correctly predicted; the model's performance is measured by its accuracy, sensitivity, and specificity. If performance is not achieved, a new model can be developed using the same algorithm and fine-tuning the algorithm's hyperparameters or another of many different machine-learning algorithms can be chosen.[3] In this study, we developed and evaluated 810 separate models before choosing the one with the highest accuracy in the validation dataset.

1.2. Concepts

Unsupervised learning: The goal is the identification of patterns that can be used to cluster observations into a preestablished number of groups. Labels are not required for this stage as algorithms will try to unbiasedly find a pattern that can structure the data. For example, clustering

algorithms are a type of unsupervised machine learning that has been applied to define intrinsic patterns of morphologic features in MDS.[4,5]

Supervised learning: In this work, the goal is the accurate assignment of the target label or classification. A model is trained with labeled examples (training dataset) and used to predict the class label of new observations. For example, machine-learning methods have been used for the prediction of cardiovascular events instead of traditional risk assessment scales commonly used in clinical practice.[6] Performance of supervised models is based on their accuracy in predicting a validation dataset, preferentially an external and independent one. Examples of supervised algorithms are random forest (an ensemble technique), decision tree, and support vector machine.

Classification model: A model that predicts discrete events, most commonly a binary target. The prediction can be "yes" or "no", or different categories, in contrast to regression models where the output is a real value.

Clustering algorithms: A method commonly used for pattern recognition within unlabeled heterogeneous datasets based on their features' similarities; applied to reduce bias, remove outliers, and separate mixed populations in high dimensional data analysis. K-means clustering is one common clustering method in machine learning. The benefit of clustering is that by portioning the data into groups that share similarities, you unbiasedly remove outliers that can introduce noise into the classification model and can increase the classification accuracy and generalization of the final classification model.

Overfitting: When a model predicts the training dataset with near 100% accuracy but fails to demonstrate the same accuracy on validation datasets, the model is likely "overfit" to the training data. This is a common problem of machine-learning models. To minimize overfitting, common methods of testing the model during training are used to identify an appropriate training stopping point.

Generalizability: When a trained model accurately labels new data, for example, the ones from different institutes or socio-economic scenarios, the model is said to generalize well. Models overfit to training data typically generalize poorly.

2. Study cohorts

A machine-learning algorithm was developed with phenotypic and molecular data from patients followed at the National Institutes of Health (NIH). The NIH cohort dataset consisted of 441 consecutive patients with any signs of marrow failure referred to the National Heart, Lung, and Blood (NHLBI) and the National Cancer Institutes (NCI). The dataset from the University of São Paulo (USP) was composed of 165 consecutive patients from the Ribeirão Preto Medical School. The training cohort included: 1) 271 consecutive patients seen at the Bone Marrow Failure (BMF) clinic at the NHLBI from 2015-2020 that had been screened for germline variants in genes related to Inherited BMF syndromes (IBMFS) by a Clinical Laboratory Improvement Amendments (CLIA)-certified targeting sequencing panel (Inherited Bone Marrow Failure Panel; University of Chicago Laboratories); 2) 18 patients from the NHLBI with pathogenic *TERT, TERC,* and *DKC1* variants identified by Sanger sequencing during 2012-2015; and 3) 112 patients from the NCI with pathogenic variants in genes related to IBMFS identified by research whole exome sequencing, targeted sequencing, and the CLIA-certified targeted panel as previously described.[7,8]

3. Genomic data curation

a. Targeting next-generation sequencing and variant calling in the USP cohort

The USP cohort was composed of 165 consecutive patients from the Medical School of Ribeirão

Preto at the University of São Paulo that were screened for germline variants in IBMFS-related genes by a customized targeting sequencing panel that was comparable to a commercial panel used for screening of the NIH cohort (supplemental Table 1). Patient's DNA was enriched with the Agilent SureSelect system and sequenced using paired-end 150-bp reads with an eight base pair sample-specific index on the Illumina technology. Peripheral blood was collected from all patients at the first clinical evaluation. Whole blood was ACK-lysed and subjected to DNA extraction with the Gentra Puregene Blood kit (Qiagen). DNA samples were kept in the laboratory and stored at -20°C. For sequencing, DNA samples were enriched with the Agilent SureSelect Target Enrichment System (Agilent Technologies) for library preparation according to manufactory instructions. Up to 96 libraries were pooled in equimolar amounts and pair-end sequenced in 300 cycles on the NextSeq platform (Illumina). A median coverage depth on targets was 350X.

Reads were aligned to the reference sequence using Burrows-Wheeler Aligner (BWA)[9] and the quality of the dataset was assessed using FastQC. Sequences were trimmed to remove adaptors as well as low-quality bases (-q 15 --minimum-length 35). Variants were called using the SAMTOOLS, DINDEL,VARDICT, and GATK pipelines.[10,11] Validity of called variants was assessed using a series of bioinformatic filters, which consider base, sequence, alignment quality metrics, a percentage of reads indicating a heterozygous variant, and any directional bias in the reads indicating a variant. Called variants were annotated using ANNOVAR.[12]

b. Systematic data analysis

For the NCI cohort, data reported had been previously curated by experts, and patients with pathogenic germline variants or variants of uncertain significance (VUS) were included in the study.

For the NHLBI cohort, results were reported for Chicago Laboratories, and per vendor, called

variants fulfilled the following criteria:

1. Mapping Quality score ≥ 20.
2. Base Quality score ≥ 10.
3. Number of variant reads > 10 reads that map to targeted regions.
4. Variants with a maximum frequency of 1% in the overall population (ExAC/gnomAD.)
5. Variants found in coding exons, affecting amino acid compositions of proteins, and variants of non-coding RNA, affecting splice sites of coding and non-coding genes.

For the USP cohort, we used an in-house pipeline designed to generate comparable data. Germline variants were called if fulfilled the following criteria:

6. Mapping Quality score ≥ 25.
7. Base Quality score ≥ 15.
8. Number of SNVs on the same read < 5.
9. Number of insertions and deletions on the same read < 2.
10. Number of total reads ≥ 20.
11. Number of variant reads ≥ 15.
12. Variants found in coding exons, affecting amino acid compositions of proteins, and variants of non-coding RNA, affecting splice sites of coding and non-coding genes.
13. Variants with a maximum frequency of 2% but > 0·1% in the overall population were manually inspected and included in the analysis according to an algorithm described in supplemental Figure 1.

Variants were filtered out if fulfilled any of the following criteria:

1. Synonymous, intragenic, intronic, and in regulatory regions. Exceptions were synonymous variants predicted to affect splicing sites, intronic regions known to harbor pathogenic

variants (such as *GATA2*), and promoter and regulatory regions known to harbor pathogenic variants (such as *TERT, TERC*, and *GATA1*).

2. Variants with maximum population frequency higher than > 2% in genome databases.
3. Variants located only in forward or reverse reads.
4. Variants not identified by visual inspection using IGV.
5. Indels of at least a four nucleotide long that were observed in multiple samples, which were considered technical artifacts.

Based on these findings, we hypothesized that we could only develop a single model for all BMF presentations if relative sizes of groups with different phenotypes were balanced in the training cohort.

c. Variant classification

Variants identified in both the NIH and USP cohorts, by either NGS assay or Sanger sequencing, were classified as pathogenic, likely pathogenic, of uncertain significance, likely benign, and benign according to the Sherloc/ACMG criteria.[13,14] In most cases, germline status was inferred if variants have been previously described at VAF around 50%, found in genes that are rarely somatically mutated, or in accordance with the patients' phenotype and disease inheritance. We also confirmed the germline status in many patients after familial investigation and segregation analysis, after a positive DEB test, or with sequencing of serial samples. In few cases only, we confirmed the germline status by sequencing a non-hematopoietic tissue.

4. Machine learning

In general, a proposed method developed in this work was a data-driven process involving feature

selection, clustering, and classification modeling. The first step was to identify the specific variables that should be included in the final model. The second step was the grouping of each cohort (an unsupervised learning step) and the last step involved the application of a classification machine-learning algorithm optimized for the Cluster A (a supervised learning step). Figure 1 shows a general pathway for the generation of a model designed for the clustering and classification of any one individual.

a. Data preparation

The NIH and USP datasets included a mix of categorical and continuous variables, including patients' ages and blood counts, and clinical parameters such as clinical manifestations, family histories, telomere lengths (TLs), and presence of abnormal karyotypes or PNH clones (Table 1). However, 4 variables were removed due to the following reasons:

Variables removed	**Reason**
White blood counts (WBC)	Highly intercorrelated with other variables
Response to immunossupresion therapy	Mapped to target
Presence of PNH clone by flow cytometry	High number of cases with missing values
Karyotype	High number of cases in the USP data with missing values

As a detection test for PNH clones is usually not requested for patients suspected of having a IBMFS, this variable was further removed from the study due to high missingness in our cohorts. Data were processed as described above and split into two groups for training, testing, and validation.

b. Feature Selection

The ReliefF method was used to rank variables by feature importance for distance-based models that use pairwise distances from the training and testing sets to maximize the predictive performance of the resulting model. Of note, several filter, wrapper, and embedded methods were also tested. It is understood that wrapper and embedded methods rely on subset evaluation and have greater potential to capture feature dependencies in predicting an endpoint, i.e. interactions.[15] Also, the ReliefF algorithm is designed to weight variable importance in datasets where a response variable is a multiclass categorical variable.[16] Finally, the basic ReliefF algorithm demonstrated the best rank order of variables based on the clinical judgment of the authors. As a result, ranking variable importance via the ReliefF algorithm was performed on the NIH training and testing data. Figures 3A and 3B show the results of feature ranking by importance via the ReliefF algorithm.

c. Clustering

Clustering was applied in an attempt to overcome the negative effects of data heterogeneity when only classification modeling was applied. Since the dataset contained a mixture of continuous and categorical data, k-prototype clustering was used. The k-prototype algorithm is similar to the k-means algorithm commonly used but more appropriate for mixed data and showed the best performance when compared to other adapted clustering methods. The clustering algorithm was used to unbiasedly split the NIH and USP datasets into clusters based on their clinical and laboratory features to generate homogeneous groups that were classified separately. Both the NIH and USP data were clustered separately but only the NIH data was used to select the ideal number of clusters to partition the data. The optimal number of clusters (k = 2) was calculated and

evaluated using the Calinski-Harabasz criterion and verified using the elbow method, two of the most common methods for determining the number of clusters within a dataset. Once the optimal number of clusters was selected, the NIH and USP data were partitioned into k = 2 groups independently using the clustering algorithm, resulting in the identification of Clusters A and B (supplemental Figure 1). Since classification models are data-hungry and the largest number of cases is always desired for machine-learning algorithms, the next steps of data modeling (a feature selection and training of the classification model) were only performed for patients assigned to Cluster A.

When deploying the method to a user app, new data will need to be assigned to an existing cluster based on the centroids established by unsupervised machine learning. To make these assignments, an Euclidean pairwise distance can be calculated between each centroid and a new data point. This can be used to identify the nearest centroid and make new data assignments. Each new data record submitted should be saved so that when significant amounts of new data are collected, the centroids may be updated.

d. Machine learning classification model selection and optimization

For this work, an ensemble-based learning and classification method was desired. Finding the minimum number of predictor variables and an optimized ensemble of trees was done using an iterative process. The optimization process compared a variety of ensemble algorithms including AdaBoost, RUSBoost, LogitBoost, GentleBoost, and Bag. Each algorithm needed to be optimized to find the best hyperparameter values within the specific algorithm's multidimensional combinations of hyperparameters. The maximum number of splits was evaluated by testing a range of integers log-scaled between 1 and one less than the number of observations. The number of learners was evaluated by testing a range of integers log-scaled between 10 and 500. A learning

rate was searched for within a range of real values log-scaled between 0·001 and 1. The number of predictors to a sample at each node was tested from within a range of 1 and the number of predictor variables. For all models and combinations of hyperparameters, an objective function was the minimum five-fold classification loss within the NIH training and testing data. For each combination, a model class was stored in a model bank for further analysis. At the conclusion, a final model was selected based on the minimization of the classification loss using the USP verification data. The USP data were not used for model training or testing. Since the target of prediction was categorical and some variables were continuous, a correlation coefficient (R) was calculated and plotted in order of variable importance (Figure 3B-C).

The optimization process included the development and evaluation of 810 separate models. Twenty-seven models were highlighted and stored in a model bank for further analysis, one model per addition of a new variable. A model that generalized the best among the 27 selected models was a bootstrapped aggregating (bag) ensemble introduced by Breiman (1996).[17] This model was originally based on the bootstrap technique of Non-parametric Statistics (Efron 1979).[17,18] The method starts with a training dataset D_t and creates new training datasets by uniform sampling with replacement from D_t to get the new datasets D_{t1},., D_f. These datasets are used to train classifiers $Y_1(x)$,...,$Y_f(x)$.[2] Within each ensemble are several individual trees that are intuitively easy to interpret. Each tree has a specified number of leaves and each leaf holds a specified number of records.[2] The leafiness of a tree is analogous to the tree's complexity. The final classification is essentially a decision by classifier committee members so that a class that occurs most often $Y_i(x)$ is selected.[2] Here, two important hyperparameters can be adjusted during the optimization process of the bagging procedure, the number of learning cycles (f), and the size of each leaf. The optimized ensemble selected based on the method described above was a bootstrap aggregation (bag) ensemble with 494 learning cycles and a minimum leaf size of four, achieving a minimum

classification loss rate of 0·0844, and an accuracy of 88·61% using 25 of the 27 features included in the training dataset.

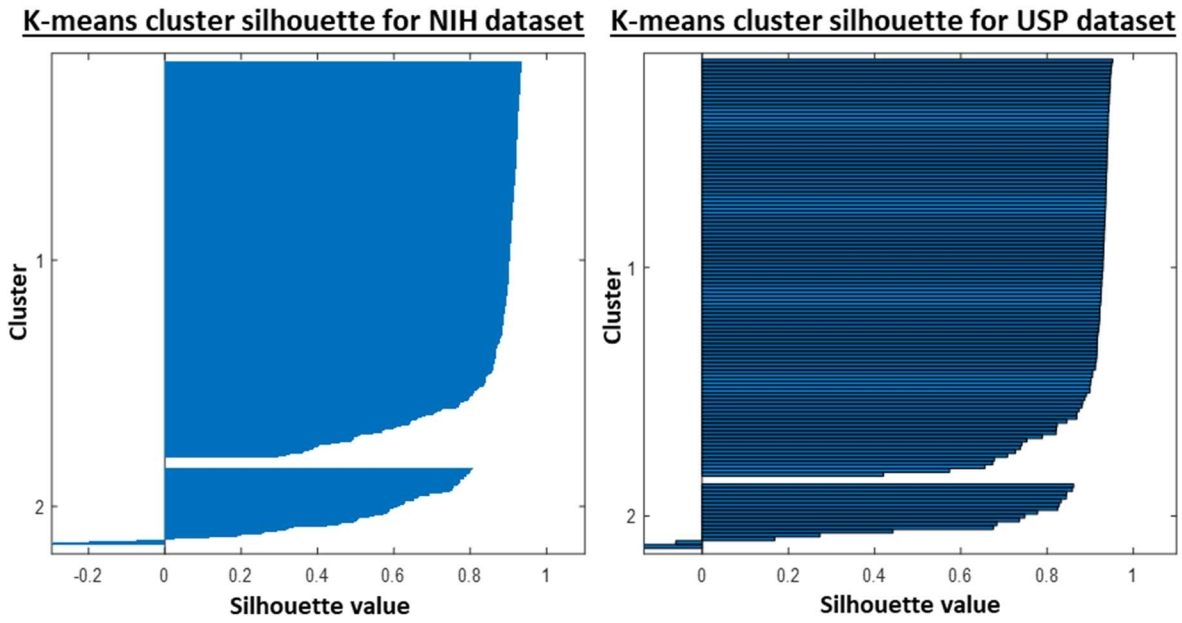
5. Logistic regression

The algorithm used to define the best cutoff (or threshold) value for binary outcome prediction by logistic regression included three steps:

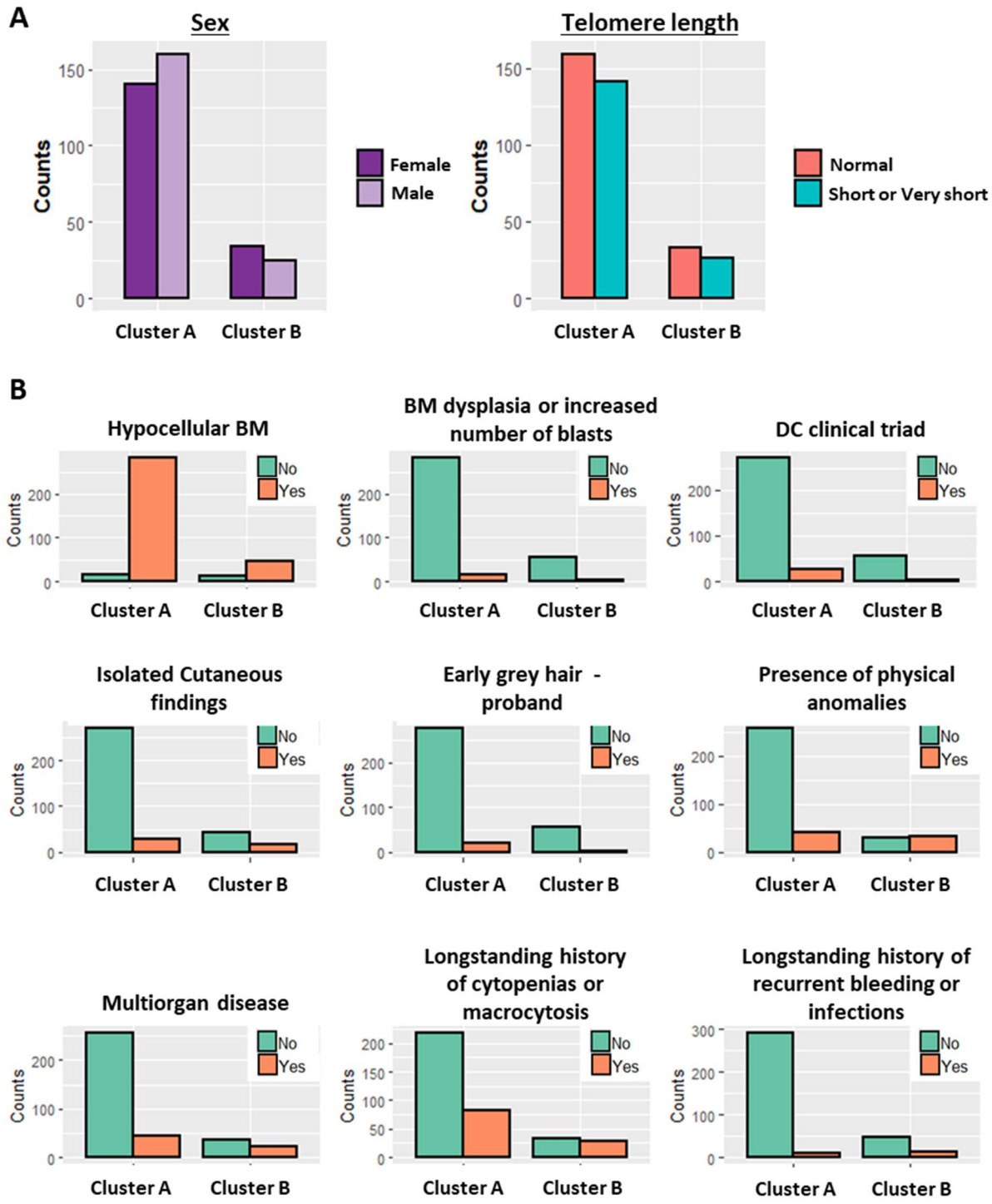
Step 1. Cutoff selection with values from 0.1 to 0.5, increased by 0.01; 41 cutoff values were tested in total.

Step 2. For each cutoff i ($i = 0.1, 0.11, \dots, 0.5$) chosen, a backward variable selection with 5-fold cross-validation in the training set was performed to find a covariate set (C_i) with the highest classification accuracy (a_i).

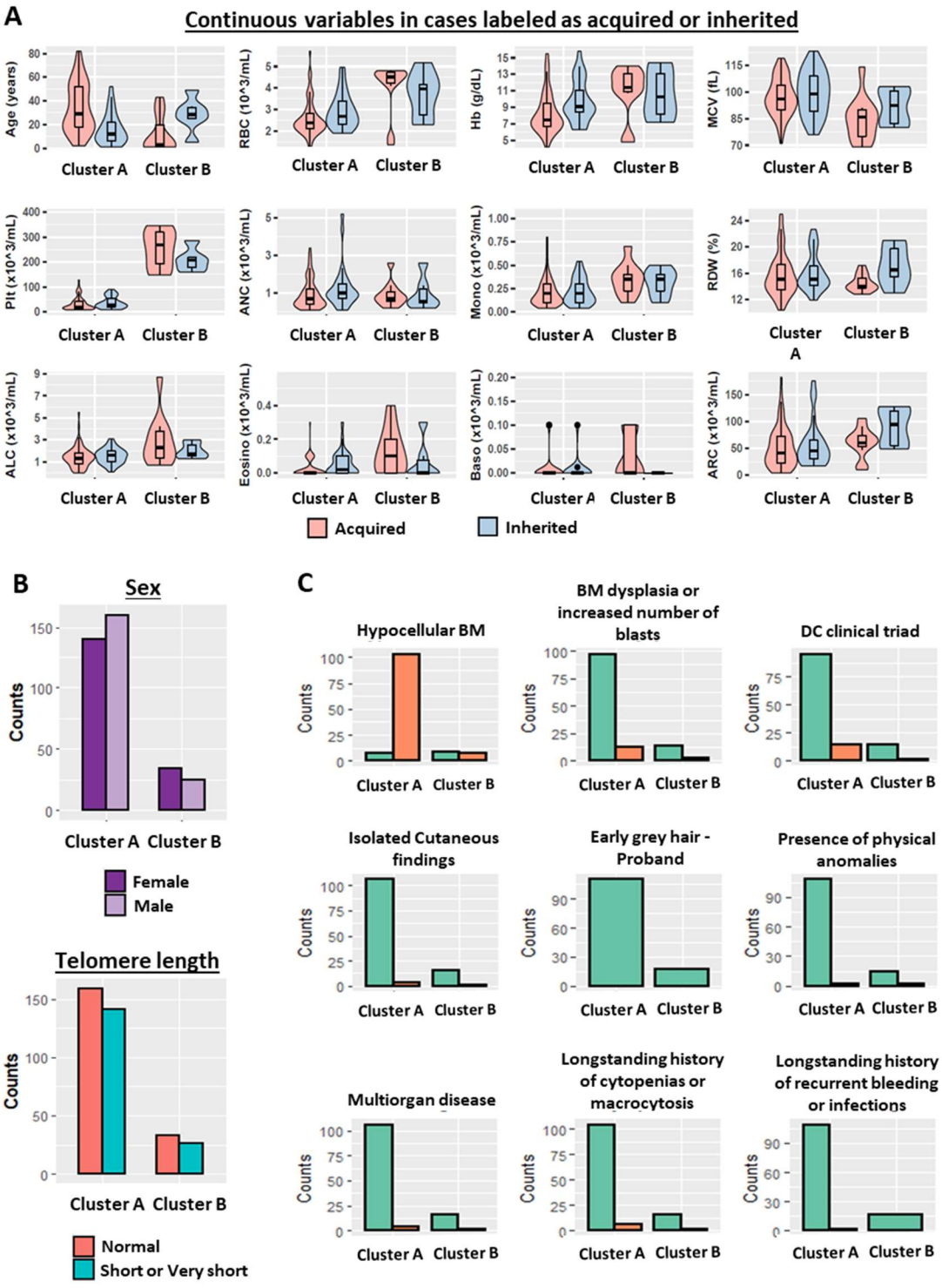
Step 3. A final model was selected based on the covariate set (C_{max}) with the highest accuracy (a_{max}.). If multiple covariate sets were obtained, a model with the shortest length of the covariate set was chosen.



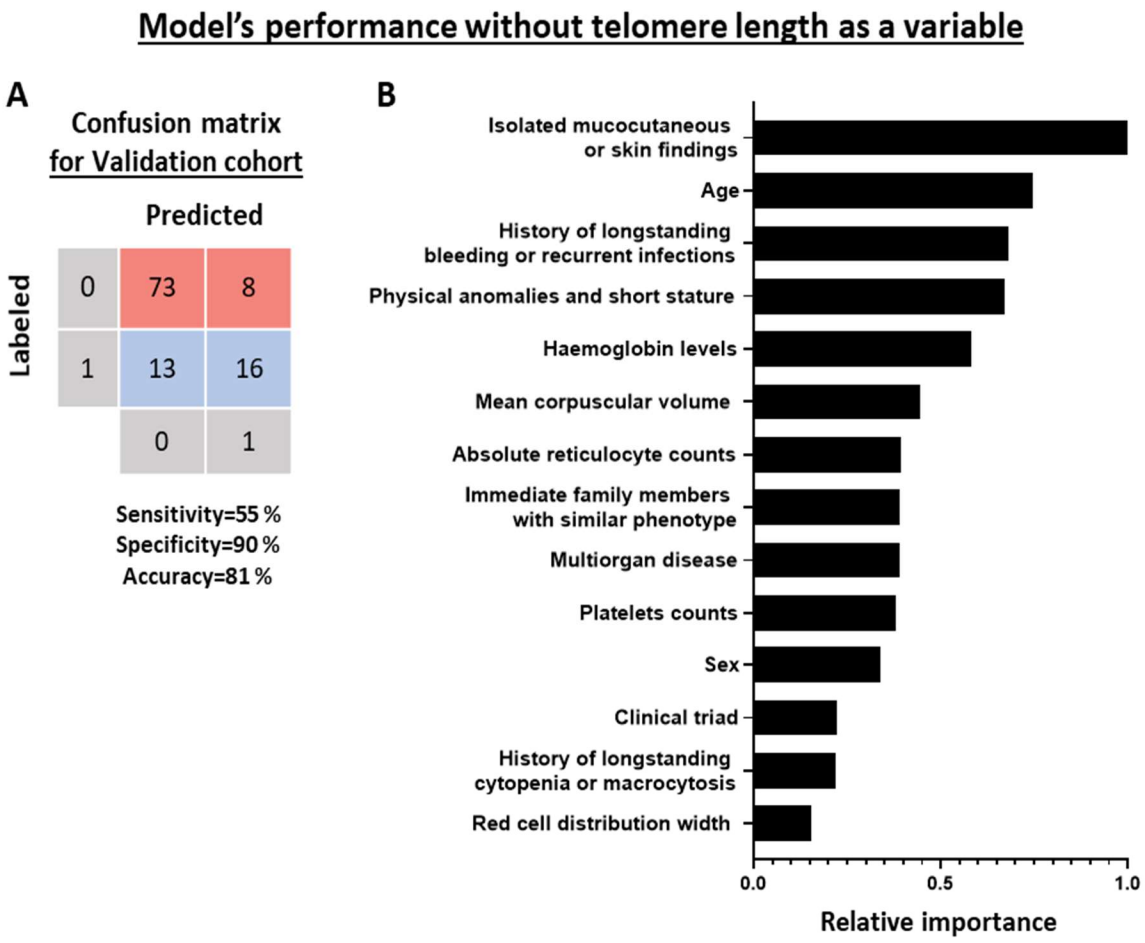
Supplemental Figure 1. Silhouette plots of clusters A and B in the training and validation datasets. Our dataset was highly dimensional. Clustering was used to partition the data into two groups. Analysis of silhouette plots gives insight as to whether these two group structures resulted in clusters that were well separated. Cluster index values plotted can range from -1 to 1. All negative values indicate a possible incorrect cluster assignment. Ranges from 0 to 1 indicate degrees of similarity of each point from within clusters. Plots show that most of the cluster index values have a high silhouette value (averages: NIH = 0.80 and USP = 0.86), indicating that Group A was somewhat separated from neighboring clusters including Group B. Furthermore, both the NIH and USP Group A silhouette plots had distinct square-like shapes. This is a good indicator that these records represent distinct groups within the data. Shapes of both Group B clusters indicated that records within Group B might not be close to the same centroid, and in both datasets, there were potentially incorrectly clustered records (3 NIH and 2 USP records). These same patterns were evident when the variants of uncertain significance data were clustered to separate the NIH and USP data. This was an indicator that these individuals should be fully tested.



Supplemental Figure 2. Baseline characteristics of categorical variables from the NIH dataset according to data clustering. (A) Bar graphs of records in Clusters A and B according to sex and telomere length measurement. (B) Bar graphs of records in Clusters A and B according to clinical variables that were included in the study.



Supplemental Figure 3. Baseline characteristics from the USP dataset according to data clustering. (A) Violin plots of continuous variables in cases labeled as acquired or inherited. (B) Bar graphs of records in Clusters A and B according to sex, telomere length measurement, and clinical variables that were included in the study.



Supplemental Figure 4. A classification model without telomere length data for predicition of bone marrow failure etiology in Cluster A . (A) A confusion matrix with prediction results for the validation cohort. Cases labeled or predicted as acquired are represented by 0 while cases labeled or predicted as inherited are represented by 1. (B) Top predictors ranked by importance by the ReliefF method.

Supplemental Table 1. Bone marrow failure-related genes screened in both the training and validation cohorts by next-generation targeted panels

Genes screened in all patients from both cohorts (n = 46)
Genes related to Fanconi anemia
BRCA2, BRIP1, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, PALB2, RAD51C, SLX4
Genes related to ribosomal diseases
RPL11, RPL35A, RPL5, RPL15, RPL26, RPS10, RPS19, RPS24, RPS26, RPS7
Genes related to telomere diseases
CTC1, DKC1, NHP2, NOP10, PARN, RTEL1, TERC, TERT, TINF2, WRAP53
Genes associated with myeloid malignancies
GATA1, GATA2, RUNX1
Severe congenital neutropenia
CSF3R, ELANE, G6PC3, GFI1, HAX1
Others
MPL, SBDS, SRP72, WAS
Genes not systematically screened in both cohorts* (n = 7)
ACD, ALAS2, BRCA1, CXCR4, DCLRE1B, DDX41, DNAJC21, EFL1, ERCC6L2, ERCC4, GRHL2, LIG4, MECOM, NAF1, PIGA, PRF1, POT1, RAD51, RBM8A, SBF2, STN1, SAMD9, SAMD9L, TP53, VPS45, UBE2T, USB1

*Not all patients were screened for mutations in these genes as they were incorporated in targeted panels at the moment they were linked to IBMFS. For patients screened for these genes, germline mutations curated as pathogenic or of uncertain significance were reported in the manuscript.

Supplemental Table 2. Detailed criteria for diagnosis of patients with bone marrow failure

Phenotype	Criteria
Isolated cytopenias	Single or bi-lineage cytopenia in peripheral blood regardless of bone marrow cellularity that not fulfill criteria for AA and MDS/hypoMDS.
Moderate aplastic anaemia (MAA)	Hypocellular bone marrow for age and at least two of the following cytopenias in peripheral blood: absolute neutrophil count < 1,500/μL, platelet count < 100,000/μL, and reticulocytes count < 60,000/μL.
Severe aplastic anaemia (SAA)	Hypocellular bone marrow for age and at least two of the following: absolute neutrophil count < 500/μL, platelet count < 20,000/μL, and reticulocytes count < 60,000/μL.
Myelodysplastic syndromes (MDS)	According to the World Health Organization (WHO) guidelines for MDS.
Hypocellular MDS (HypoMDS)	Meeting WHO criteria for MDS with marrow cellularity of ≤ 25%.
Dyskeratosis congenita (DC)	At least two of three manifestations of the clinical triad (dystrophic nails, patchy skin hyperpigmentation, and oral leukoplakia) and telomere length below the 1st percentile for age-matched controls.
Fanconi anaemia (FA)	Biallelic mutation in a known FANC gene and/or positive chromosome breakage analysis in lymphocytes and/or skin fibroblasts.
Shwachman Diamond syndrome (SDS)	- Fulfill the combined presence of haematological cytopenia of any given lineage (most often neutropenia) and exocrine pancreas dysfunction. - Hematologic abnormalities may include: a. Neutropenia < 1.5 x 10^9/L on at least two occasions over at least 3 months. b. Hypoproductive cytopenia detected on two occasions over at least 3 months. - Tests that support the diagnosis but require corroboration: a. Persistent elevation of hemoglobin F (on at least two occasions over at least 3 months apart). b. Persistent red blood cell macrocytosis (on at least two occasions over at least 3 months apart), not caused by other etiologies such as hemolysis or a nutritional deficiency. - Pancreatic dysfunction may be diagnosed by the following: a. Reduced levels of pancreatic enzymes adjusted to age [fecal elastase, serum trypsinogen, serum (iso)amylase, and serum lipase]. - Fulfill the combined presence of hematological cytopenia of any given lineage (most often neutropenia) and exocrine pancreas dysfunction. - Hematologic abnormalities may include: a. Neutropenia < 1.5 x 10^9/L on at least two occasions over at least 3 months. b. Hypoproductive cytopenia detected on two occasions over at least 3 months. - Pancreatic dysfunction may be diagnosed by the following: a. Reduced levels of pancreatic enzymes adjusted to age [fecal elastase, serum try*p*sinogen, serum (iso)amylase, and sand erum lipase].
Diamond Blackfan syndrome (DBA)	- Clinically: Anemia presenting on or before the third year of life with reticulocytopenia and greatly reduced or absent bone marrow erythroid precursors. - Genetically: The presence of a mutation of disease-associated gene in combination with clinical characteristics of DBA.

Supplemental Table 3. Categorical variables included in the study: clinical manifestations, family histories, and laboratory tests

Variable	Categories	Description
Telomere length measurement	• TL<1st percentile of age-matched controls • TL<10th percentile of age-matched controls • Normal TL (>10th percentile)	TL was preferentially measured by Flow-FISH. In the NIH cohort, TL from lymphocytes was recorded for each patient. In the USP cohort, TL from total leukocytes was used in the analysis. For NIH patients seen at the institution between 2011 and 2015, TL was measured by either Southern blot or quantitative PCR in total leukocytes. TL below the 1st percentile of age-matched controls was considered very short and below the 10th percentile were considered short. Patients with TL above the 10th percentile were deemed to have normal TL.
Bone marrow cellularity for age	• Hypocellular BM for age • Normo or hypercellular BM for age	Cellularity assessed by BM biopsy.
Bone marrow dysplasia	• Yes • No	Presence of bone marrow dysplasia that fulfills criteria for MDS according to the WHO 2016 guidelines. Increased blasts defined as > 5%.
Presence of the DC clinical triad	• Yes • No	At least two of three manifestations of the clinical mucocutaneous triad associated with dyskeratosis congenita (dystrophic nails, skin hyper- or hypo-pigmentation, and oral leukoplakia).
Isolated mucocutaneous or skin findings	• Yes • No	Presence of any of these: skin hyper- or hypo-pigmentation, leukoplakia, nail dystrophy, cafe au lait spots, petechiae, skin infections, skin rash or lesions, ichthyosis, mouth ulcers and macules, nevi, and warts.
Early hair greying	• Yes • No	Probands with full or patchy hair greying at a young age, usually below 30 yo.
Physical anomalies and short stature	• Yes • No	Presence of any of these: short stature, microcephaly, and thumb and arm anomalies (absent, hypoplastic, duplicated, or flat radius, thumbs, thenar, and fingers), clinodactyly, polydactyly, syndactyly, Fanconi face and dysmorphic features, palate high arch, skeletal anomalies (Klippel-Feil and Sprengel deformities), renal anomalies (horseshoe, hypoplastic, and others), gonads anomalies, and failure to thrive.
Multiorgan disease	• Yes • No	Presence of any of these: - Liver diseases: fatty liver, cirrhosis, steatosis, splenomegaly, portal hypertension, nodular regenerative hyperplasia, and NASH. - Pulmonary diseases: pulmonary fibrosis, interstitial pneumonitis, chronic obstructive disease, and clubbing. - GI dysfunction: chronic diarrhea, atresia, pancreatic insufficiency, esophageal dilatations or strictures, epistaxis, dysphagia, colitis, and malabsorption. - Neurological findings: cerebellar ataxia, growth and development delay, cerebral palsy, brain calcifications, and cysts. - Others: Ear infections or hearing loss, septo-optic dysplasia, and retinopathies. - Congenital cataracts, Coats plus syndrome and hip necrosis.

History of longstanding cytopenia or macrocytosis	• Yes • No	Individuals with a history of cytopenias or macrocytosis: - At birth - Since childhood - For more than 5 years
History of longstanding bleeding or recurrent infections	• Yes • No	Individuals with history of recurrent bleeding or infections during childhood or adult life: - During childhood - For more than 5 years
Immunodeficiency	• Yes • No	Presence of any of these: - Common variable immunodeficiency - Hypogammaglobulinemia - B and NK lymphopenia - T lymphopenia
Immediate family members with similar phenotypes	• Yes • No	Any parent, sibling, or offspring with hematologic findings from the spectrum of IBMFS (including AML and MDS), multiorgan disease, or physical abnormalities linked to proband's phenotypes.
Extended family members with similar phenotypes	• Yes • No	Any relative other than immediate family members with hematologic findings from the spectrum of IBMFS (including AML and MDS), multiorgan disease or physical abnormalities linked to proband's phenotypes.
Relatives with early hair greying	• Yes • No	Individuals whose relatives of any degree had full or patchy hair greying at a young age, usually below 30 yo.

Supplemental Table 4. Logistic regression analysis for prediction of bone marrow failure etiology

	Univariable logistic model		Multivariable logistic model	
	OR (95%CI)	***P*-value**	**OR (95%CI)**	***P*-value**
Telomere length: <1st vs normal	57.45 (25.78, 144.14)	**< .001**	16.71 (4.18, 89.68)	**< .001**
Telomere length: <10th vs normal	6.29 (2.43, 17.16.	**< .001**	2.87 (0.57, 16.88)	.210
Age	0.95 (0.93, 0.97)	**< .001**	0.93 (0.88, 0.97)	**.001**
Presence of physical anomalies and short stature	0.05 (0.02, 0.11)	**< .001**	0.09 (0.02, 0.46)	**.005**
Dyskeratosis congenita mucocutaneos triad	0.01 (0, 0.05)	**< .001**	0.04 (0, 0.32)	**.009**
Abnormal cutaneous findings	0.06 (0.02, 0.5)	**< .001**	0.03 (0, 0.16)	**< .001**
Presence of multiorgan diseases	0.04 (0.01, 0.09)	**< .001**	0.04 (0.01, 0.17)	**< .001**
Eosinophils counts	4.29 (2.22, 9.01)	**< .001**	4.02 (1.21, 17)	**.039**
Mean corpuscular volume	1.06 (1.04, 1.09)	**< .001**	1.04 (0.99, 1.1)	.094
Basophil counts	16.77 (3.58, 94.24)	**< .001**	1.8 (0.05, 29.91)	.730
Red blood cell counts	2.01 (1.41, 2.9)	**< .001**		
Monocyte counts	1.92 (1.61, 2.33)	**< .0011**		
Neutrophil counts	1.68 (1.31, 2.21)	**< .001**		
Haemoglobin levels	1.5 (1.32, 1.72)	**< .001**		
Platelet counts	1.03 (1.02, 1.04)	**< .001**		
Reticulocyte counts	1.02 (1.02, 1.03)	**< .001**		
History of longstanding cytopenias or macrocytosis	0.21 (0.12, 0.37)	**< .001**		
Immediate family members with similar phenotypes	0·15 (0·08, 0.26)	**< .001**		
Extended family members with similar phenotypes	0.43 (0.23, 0.8)	**.008**		
Sex	2.05 (1.24, 3.46)	**.006**		
Early grey hair	0.23 (0.09, 0.58)	**.002**	0.16 (0.03, 0.86)	**.034**
Bone marrow cellularity: Normocellular vs hypocellular	6.38 (2.07, 23.78)	**.002**		
Immunodeficiency	0.96 (0.3, 3.63)	.951	23.44 (2.12, 342.08)	**.013**
Red cell distribution width	0·99 (0.91, 1.07)	.822		
Lymphocyte counts	0.95 (0.7, 1.28)	.746		
Relatives with early grey hair	0.63 (0.29, 1.45)	.263		
Longstanding history of recurrent bleeding and infections	0.33 (0.08, 1.28)	.105		
Bone marrow dysplasia consistent with MDS or increased number of blasts	0.41 (0.14, 1.14)	.081	0.18 (0.02, 1.41)	.099

References

1. Shouval R, Fein JA, Savani B, Mohty M, Nagler A. Machine learning and artificial intelligence in haematology. *Br J Haematol*. Jan 2021;192(2):239-250. doi:10.1111/bjh.16915
2. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. Jun 2015;16(6):321-32. doi:10.1038/nrg3920
3. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 12 21 2019;19(1):281. doi:10.1186/s12911-019-1004-8
4. Brück OE, Lallukka-Brück SE, Hohtari HR, et al. Machine Learning of Bone Marrow Histopathology Identifies Genetic and Clinical Determinants in Patients with MDS. *Blood Cancer Discov*. May 2021;2(3):238-249. doi:10.1158/2643-3230.BCD-20-0162
5. Nagata Y, Zhao R, Awada H, et al. Machine learning demonstrates that somatic mutations imprint invariant morphologic features in myelodysplastic syndromes. *Blood*. 11 2020;136(20):2249-2262. doi:10.1182/blood.2020005488
6. Munger E, Hickey JW, Dey AK, Jafri MS, Kinser JM, Mehta NN. Application of machine learning in understanding atherosclerosis: Emerging insights. *APL Bioeng*. Mar 2021;5(1):011505. doi:10.1063/5.0028986
7. Ballew BJ, Yeager M, Jacobs K, et al. Germline mutations of regulator of telomere elongation helicase 1, RTEL1, in Dyskeratosis congenita. *Hum Genet*. Apr 2013;132(4):473-80. doi:10.1007/s00439-013-1265-8
8. Alter BP, Giri N, Savage SA, Rosenberg PS. Cancer in the National Cancer Institute inherited bone marrow failure syndrome cohort after fifteen years of follow-up. *Haematologica*. 01 2018;103(1):30-39. doi:10.3324/haematol.2017.178111
9. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. Jul 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324
10. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. Sep 2010;20(9):1297-303. doi:10.1101/gr.107524.110
11. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 06 2016;44(11):e108. doi:10.1093/nar/gkw227
12. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. Sep 2010;38(16):e164. doi:10.1093/nar/gkq603
13. Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med*. 10 2017;19(10):1105-1117. doi:10.1038/gim.2017.37
14. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. May 2015;17(5):405-24. doi:10.1038/gim.2015.30
15. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. A review of feature selection methods on synthetic data.
16. Robnik-Šikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF.
17. Breiman L. Bagging predictors.
18. Efron B. Computers and the theory
of statistics: thinking the unthinkable. *SIAM Rev* . 1979. p. 460-480.