Blood count phenotype and blood cancer risk prediction

Imperial University MSc Health Data Analytics and Machine Learning

Cameron Appel Kate Cheng David Ensor

09/05/2023

Abstract

Blood cancer affects over 250,000 people in UK and early diagnosis is essential for early intervention and treatment to improve patient outcomes and reduce mortality. Blood tests are important diagnostic tests for multiple blood cancer types and blood counts are demonstrated to be predictive of cancer progress in other types of cancers. This study uses UK Biobank data on blood count and demographic and behavioural cofactors to develop models for blood cancer risk prediction. Stability selection was employed for variable selection, followed by fitting Cox proportional hazards models and logistic regression models for prediction. Tree based models such as random forest and XGBoost were also explored. Clustering has been applied to look for potential blood count profiles that are predictive of blood cancer risk. The methods are compared using C-index for time-to-diagnosis outcome and AUC for binary outcomes.

This study found that red blood cell distribution width and platelet distribution width to be important factors for blood cancer risk prediction. These results are in coherent with previous results for other types of cancer. These could be incorporated into early diagnosis procedures. The blood counts found to be important in prediction can be used as early diagnosis indicators.

Introduction

Over 40,000 people are diagnosed with blood cancer each year in the UK, and over 250,000 people are currently living with blood cancer. (Blood Cancer UK n.d.). It is more common in adults age 60 and above but can also affect children, adolescents and young adults (Kucine 2020). Early diagnosis is key to improving health outcomes and reducing mortality.

Blood cancers encompasses a broad range of cancers that are classified by the type of blood cell type that is affected in the mutation (Arber et al. 2016). Generally the disease is progressive and resulting in changes in cell types with increases or decreases in concentration of blood cells types depending on the blood cancer type and corresponding blood cell lineage affected. Blood cancers can be caused by genetic abnormalities causing abnormal changes in the formation of blood cells. Certain blood cancers such as myeloprolifeative neoplasts such as polycythaemia vera, essential thrombocythemia and myelofibrosis have been associated with mutations in the JAK2, CALR and MPL genes (Grinfeld et al. 2018) or FLT3 is most frequently detected in patients with AML (Döhner, Wei, and Löwenberg 2021).

Using data for prediction of cancer has been a long term goal in medicine, back in 2014 one study demonstrated success in predicting leukaemia in patients based on the cell population data taken from full blood counts(FBC) using excel spreadsheet methods (Yang et al. 2014). Machine learning (ML) has improved performance. Previous studies have performed similar attempts with good results assessing specific blood cancers such as leukemia as well as more broad classifications with significant improvements in accuracy(M. El-Halees and H. Shurrab 2017).

This method is not unique to blood cancer but has been applied to other health conditions and it was demonstrated that higher platelet ratio is associated increased in hospital mortaility in patients with myocardial infarction (Yao et al. 2023).

There are other approaches to use machine learning in diagnosis such as convoluted neural networks to identify abnormal leucocytes on histology slides with significant success however misclassification errors still exist especially between cell types that have similar lineage (Shahid and Singh 2019; Wang et al. 2019).

Although there is literature available demonstrating prediction using blood counts for specific blood cancer types there has not been any studies aiming for prediction across a broad category of blood cancers. Our aim is to investigate if it was possible to identify early changes in blood cell concentrations present prior to diagnosis of blood cancer regardless of type in order to predict cancer. This could potentially identify an early cancer biomarker which could improve prediction of blood cancer diagnosis and allow for earlier intervention.

We aimed to apply supervised and unsupervised machine learning techniques to identify blood biomarkers that could predict blood cancer and sub groups of blood cancer. This could potentially develop into a

predictive model to assess for variables that could be a catalyst for research as well as improvement in clinical assessment to improve diagnosis, prognosis and treatment of myeloproliferative disease.

Methods

Data processing

UK biobank data was reviewed for patients with diagnosis of blood cancers according to ICD 10 and ICD 9 coding. We merged the ICD 9 and ICD 10 groups which had corresponding diagnosis according to the WHO classification for blood cancer (Arber et al. 2016). If there was not a corresponding ICD10 code for the ICD9 code then this was added to the Other cancer group. Our controls were retrieved from the UK Biobank and were identified as healthy participants with no ICD 10 & 9 diagnosis. We further grouped the blood groups according to the cell histology resulting in four subgroups classified as per the cell line: lymphoid, Myeloid, Dendritic and histiocytic cell and Other cancers. Figure 1 illustrates the cell lineage of a hematopoietic cell.

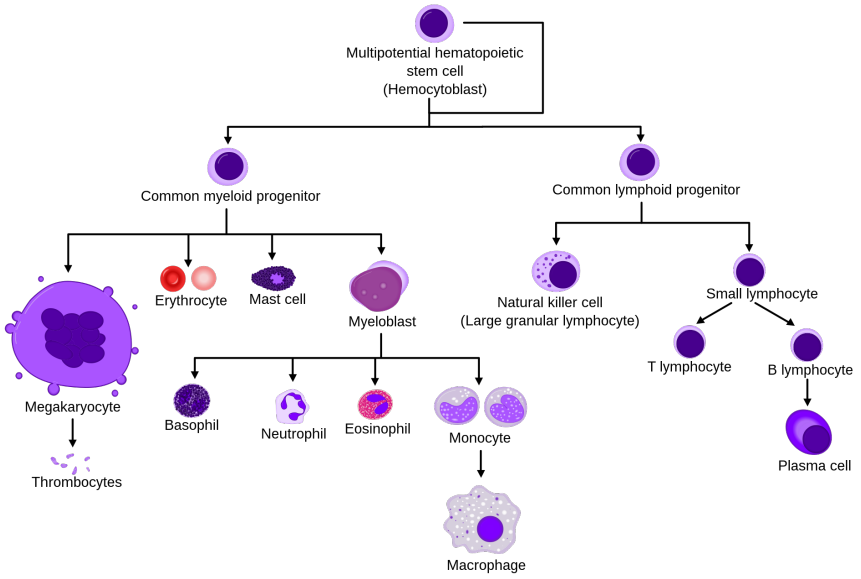


Figure 1. Cell Lineage of a Hematopoietic Cell

Many FBC components are mathematically related and so can be derived from each other. In order to prevent correlation and to ensure variables are independent we calculated the correlation between each pair of variables and removed correlated blood components.

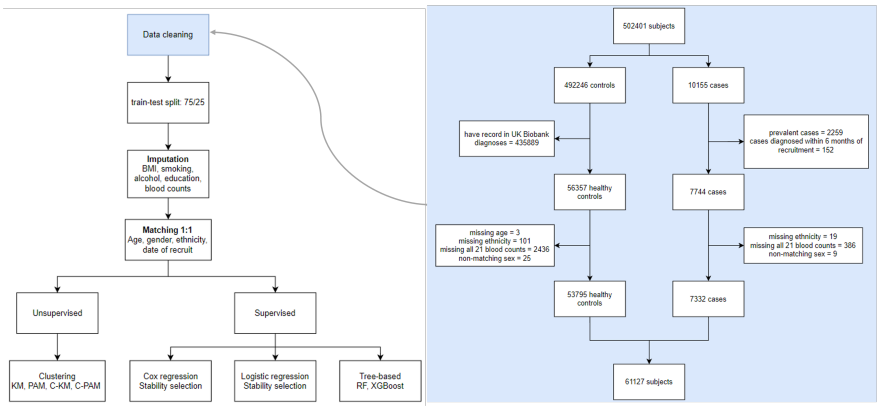


Figure 2. Data cleaning process.

We further grouped the blood groups according to the cell histology resulting in four subgroups classified as per the cell line: lymphoid, myeloid, dendritic and histiocytic cell and other cancers. The Dendritic and Histiocytic cell group had a low number and it was not large enough to be assessed so we removed it.

After data cleaning, the dataset is split into train (75%) and test (25%) set. k-nearest neighbour imputation is used to impute missing values separately in train and test set. Cases and controls are matched to 1:1 on age, sex and date of recruitment in the train set.

Stability selection

Stability selection on Cox regression (blood cancer case-control status) and logistic regression (blood cancer subtype case-control status) was performed using the sharp package in R (Bodinier et al. 2021). All covariates and blood count features were supplied to the models, the penalty factor argument was set to 0 for all covariates to keep them in all models and to 1 for blood count features to select a subset of blood count features. Two parameters, penalty parameter (λ) and selection proportion (π), are tuned by subsampling and maximising the stability score. A grid of parameter values was defined using the default setting in the `VariableSelection` function, which uses a sequence from 0.6 to 0.9 with 0.01 interval for π and `LambdaGridRegression` function for generating λ values. For each pair of parameter values, 100 models were built using subsamples from the training data. The stability score was defined using a likelihood function that measures the deviation from an unstable model that uniformly selects features and the most stable models would stably select and exclude blood count features across 100 iterations.

After stability selection, unpenalised Cox regression and logistic regression models were fit using all covariates and the selected blood counts. The models are used to predict outcome in the test set and compared to the true outcomes. C-index and AUC are calculated for time-to-event and binary outcomes, respectively.

XGBoost

We applied XGBoost (XGB) algorithm to all cases and each subgroup (Chen and Guestrin 2016) . Each subgroup algorithm was trained and tuned using the caret package separately (Kuhn 2008). The models were tuned with repeated cross validation number 10 and 3 repeats. Cross validation was set at 10 times with a tuning of nrounds (500, 1000, 1500) with a max depth (2,4,6) eta (0.2), gamme (0), col sample_by tree (1) minchild weight (1) subsample (1). The default loss function was used:'binary:logistic.' This was applied to all cancer groups and the optimal settings were selected based on highest accuracy.

Random Forest

Random Forest (RF) package was used (Liaw and Wiener 2002). Each model was trained with repeated cross validation number 10 and 3 repeats tuning at 250,500,1000,1500 trees to find the optimal number of trees and m try (1 to 7) and optimal mtry and ntree was selected by highest optimal OOB accuracy and computational efficiency for each final random forest model for each cancer group.

Clustering

Cluster analysis was performed on blood counts to explore potential groups of observations that are associated with blood cancers, as well as to reduce dimensionality for further analysis. After clusters were found, a logistic regression of cluster membership against blood counts was performed to characterise the groups and determine which variables determine the cluster membership. Then, logistic regression model of blood cancer outcome against clusters and covariate were also fitted. This model was then used to predict the outcome in the test set. The predictions are compared to true outcome and AUC were calculated. These models were then compared with the performance of regression models resulting from other models.

Clustering Cluster analysis was performed on blood counts to explore potential groups of observations that are associated with blood cancers, as well as to reduce dimensionality for further analysis. Multiple clustering methods and distance metrics were explored, K-Means clustering was implemented using sharp package in R (Bodinier et al. 2021), pam was done using the package cluster (ref) and consensus clustering is done using the package ConsensusClusterPlus.

For each method, once cluster membership had been determined, a 70-30 train-test split was applied to the dataset. A logistic regression model of blood cancer outcome against clusters and covariates was then fitted on the training dataset. This model was then used to predict the outcome of each participant in the testing dataset. The ROC and AUC of these predictions were then calculated and compared with the performance of logistic regressions using just blood counts and covariates. For KMeans clustering, the number of clusters must be determined a priori. The Elbow and Silhouette methods were used, and the best number of clusters was identified based on both. Consensus clustering was performed based on K-Means or PAM clustering. The best number of clusters is defined using the change in area under consensus CDF. For simlpicity in presentation, consensus K-Means and PAM would be refered to as C-KM and C-PAM in the rest of the report.

Results

Table 1. Demographics of the matched dataset.

	Control	Cases
n	5485	5485
Sex = Male (%)	3086 (56.3)	3017 (55.0)
Age at recruitment (median[IQR])	62.00 [57.00, 66.00]	62.00 [57.00, 66.00]
BMI (median[IQR])	26.21 [23.90, 28.89]	27.13 [24.56, 30.36]
Ethnicity (%)		
African	22 (0.4)	32 (0.6)
Asian	62 (1.1)	89 (1.6)
Caribbean	53 (1.0)	61 (1.1)
White	5295 (96.5)	5219 (95.2)
Mixed	15 (0.3)	18 (0.3)
Other	18 (0.3)	40 (0.7)
Prefer not to answer	20 (0.4)	26 (0.5)

	Control	Cases
Smoking status (%)		
Never	2917 (53.2)	2940 (53.6)
Previous	2066 (37.7)	2030 (37.0)
Current	502 (9.2)	515 (9.4)
Alcohol frequency (%)		
Never	384 (7.0)	407 (7.4)
Rarely	630 (11.5)	612 (11.2)
Monthly	550 (10.0)	545 (9.9)
1-2 times per week	1380 (25.2)	1354 (24.7)
3-4 times per week	1264 (23.0)	1234 (22.5)
Daily	1277 (23.3)	1333 (24.3)
Highest education (%)		
A level	155 (2.8)	172 (3.1)
College	668 (12.2)	648 (11.8)
CSE	239 (4.4)	238 (4.3)
GCSE	993 (18.1)	1100 (20.1)
NVQ	739 (13.5)	629 (11.5)
Other professional	1589 (29.0)	1605 (29.3)
Others	1102 (20.1)	1093 (19.9)

Comparison of model prediction accuracies

Table 2. Comparison of Cox regression models using stably selected blood count features and using covariates only.

Model	Variables	C.index
Blood cancer vs control	Covariates	0.61
Stability selection	Covariates + blood counts	0.66
PAM	Covariates + cluster	0.63
C-KM	Covariates + cluster	0.63

Table 3. Comparison of classification models. The model evaluation metric used is AUC. The outcomes are binary case-control statuses for pooled blood cancer types or subtypes. The 'Covariates' models are logistic regression models containing only the covariates. The 'Stability selection' models are logistic regression models containing covariates and selected blood count features. The clustering models are logistic regression models containing covariates and the cluster membership variables.

Outcome	Covariates	Stability_selection	RF	XGBoost	PAM	C-KM
Blood cancer	0.60	0.66	0.70	0.69	0.63	0.62
Lymphoid type	0.59	0.63	0.65	0.59	0.59	0.58
Myeloid type	0.67	0.72	0.75	0.70	0.69	0.69
Other	0.63	0.75	0.75	0.71	0.69	0.68

Overall, including blood count features in modeling increases the accuracy of the models. Random forest models have the best performance for all outcomes.

Stability selection

Cox regression Seven blood count features are stably selected for the time to blood cancer diagnosis outcome. In the unpenalised model, immature reticulocyte fraction was removed due to having extremely large hazard ratio. The removal did not affect the model accuracy. For the other variables, higher red blood cell count is associated with lower hazard of blood cancer diagnosis, whereas red blood cell distribution width and platelet distribution width are found to associated with higher hazard (Fig. 2).

Logistic regression For the blood cancer subtypes, seven blood count features were selected for lymphoid, four for myeloid and six for other types of blood cancers. Among the selected features, red blood cell distribution width was selected for all subtypes.Platelet distribution width and mean corpuscular volume are selected for both myeloid and other types. Red blood cell count and white blood cell count are selected for both lymphoid and other types (Fig. 3).

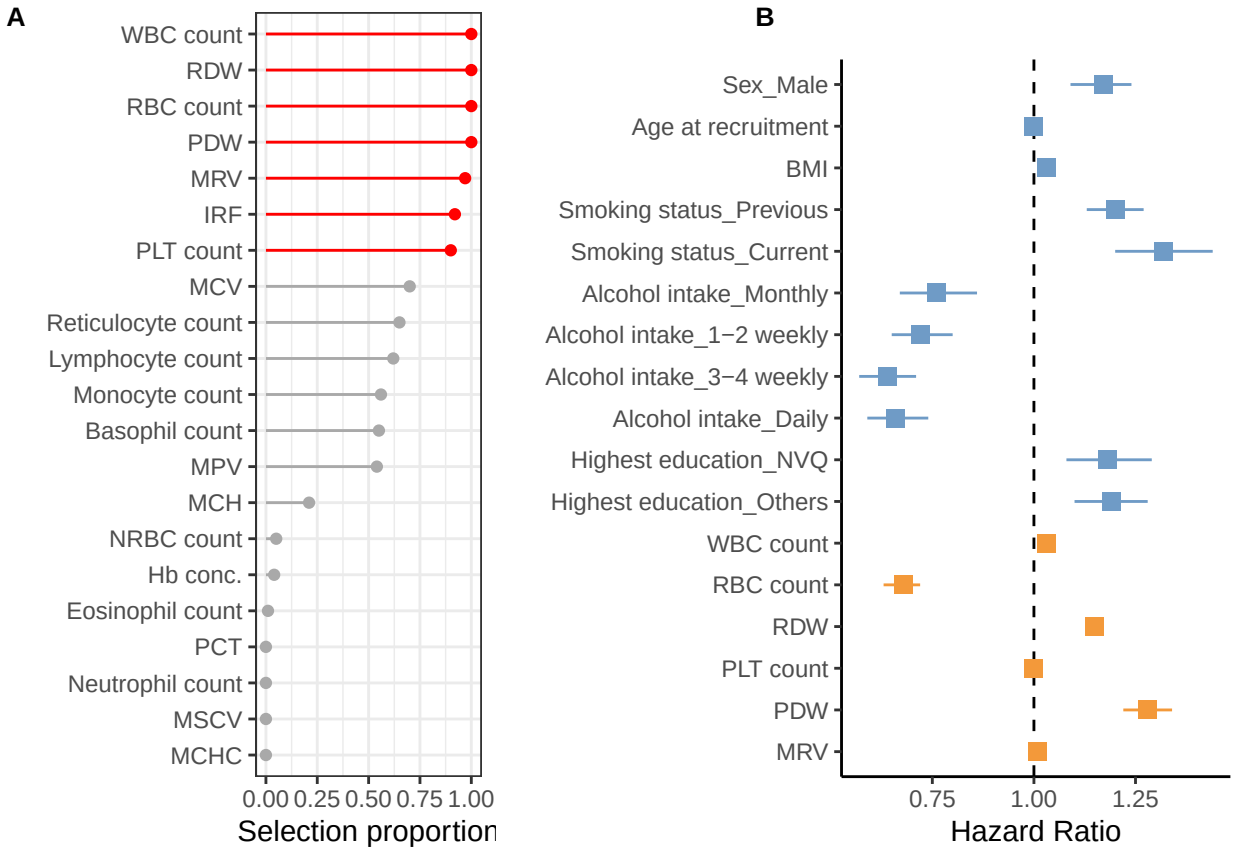


Figure 3. Results for Cox regression model. A shows the selection proportion for each blood count feature, the stably selected features are labeled in red. B shows the hazard ratios for an unpenalised Cox regression model fitted using selected blood counts (orange) and all covariates (blue).

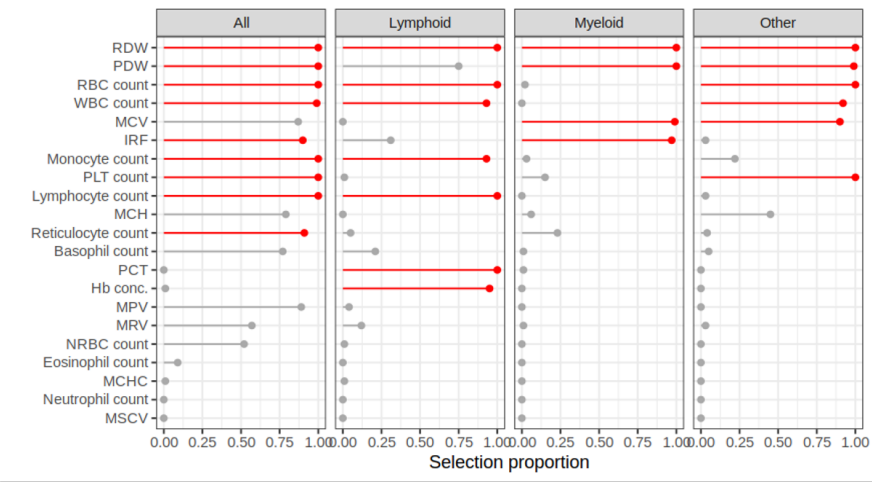


Figure 4. Stability selection results for logistic regression against blood cancer subtypes.

The unpenalised logistic regression models found the selected blood counts to be significantly associated with the outcomes, except for white blood cell count for lymphoid type, which is marginally significant ($p = 0.055$). Red blood cell distribution width is found to associated with higher odds of developing all the subtypes, consistent with the result found for blood cancer in general. Platelet distribution width and mean corpuscular volume are found to increase the odds of myeloid and other types of blooc cancer, also consistent with the results for pooled blood cancer. Red blood cell count was found to associate with lower odds of lymphoid and other types of blood cancer, again consistent with the pooled result. For lymphoid type cancer, platelet crit was found to associate with lower odds and the count of lymphocyte and monocyte, two types of white blood cells, are found to associate with high odds of lymhpoid type blood cancer. Ir fraction is again removed from the unpenalised model of myeloid type blood cancer due to an extremely high OR. The removal improved the model accuracy marginally from AUC = 0.69 to AUC = 0.70.

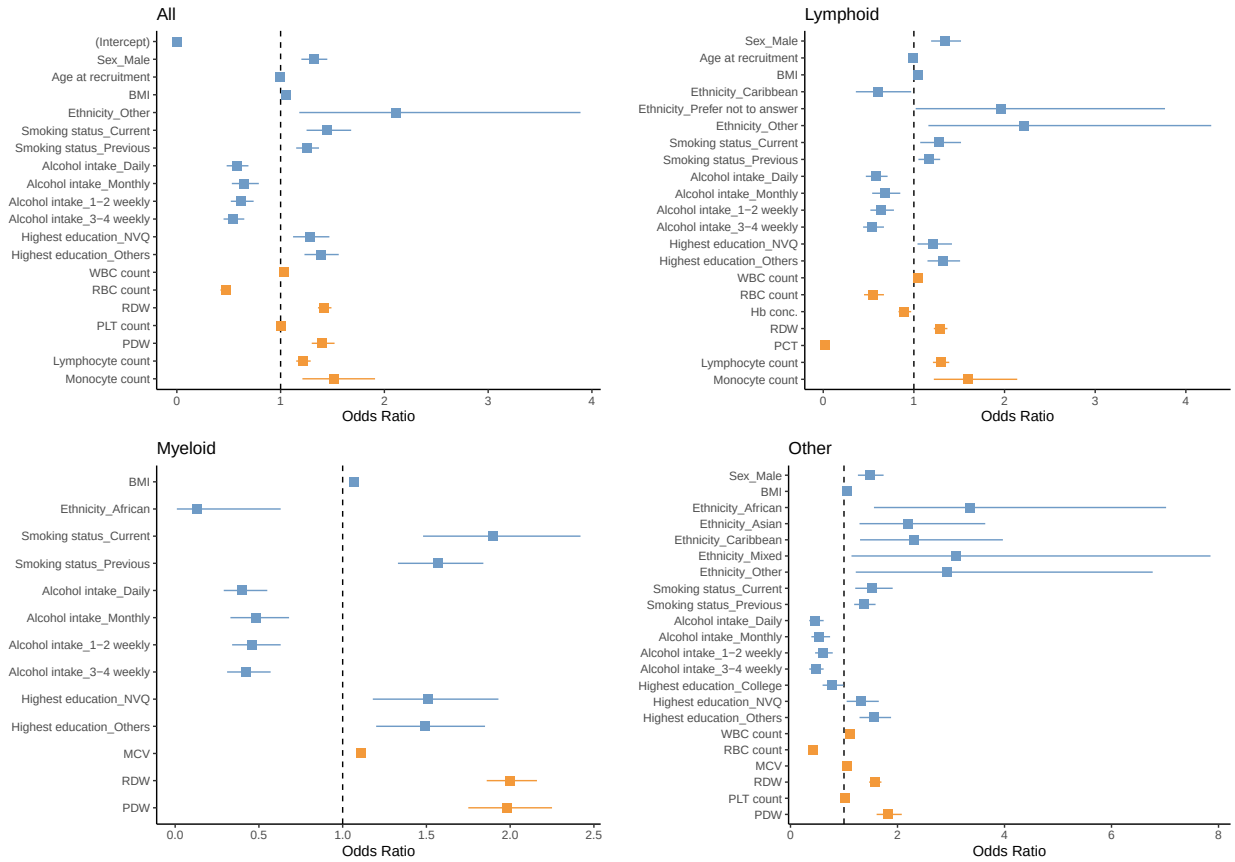


Figure 5. Unpenalised logistic regression results for blood cancer subtypes.

Tree-based methods

The AUC values demonstrate that the random forest models performed better than XGBoost across all cancer and subgroups. The prediction accuracy for XGB models of all subtypes were similar except for lymphoid type.

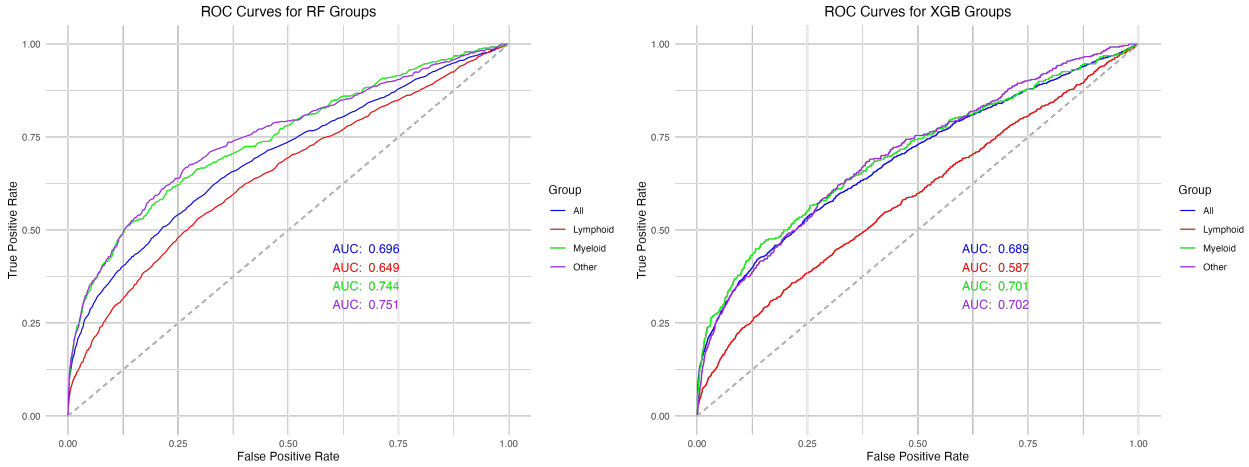


Figure 6. ROC curves and AUC for random forest and XGBoost models

In order to compare feature importance for categorical prediction of both models and we selected the gain metric for XGB models and Mean Decrease Gini metric in RF as these are similar indicators of which features used in splitting decision trees. These results are scaled and cut off was set at 80% variable importance. The Mean decrease in Gini coefficient measures how each variable contributes to the homogeneity of the node and leaves in the resulting RF. The higher the importance of the variable, the higher the value of the mean decrease gini score. The gain metric indicates the relative contribution of the feature by assessing the feature contribution to each tree in the XGB model - the higher the variable importance, the higher the contribution.

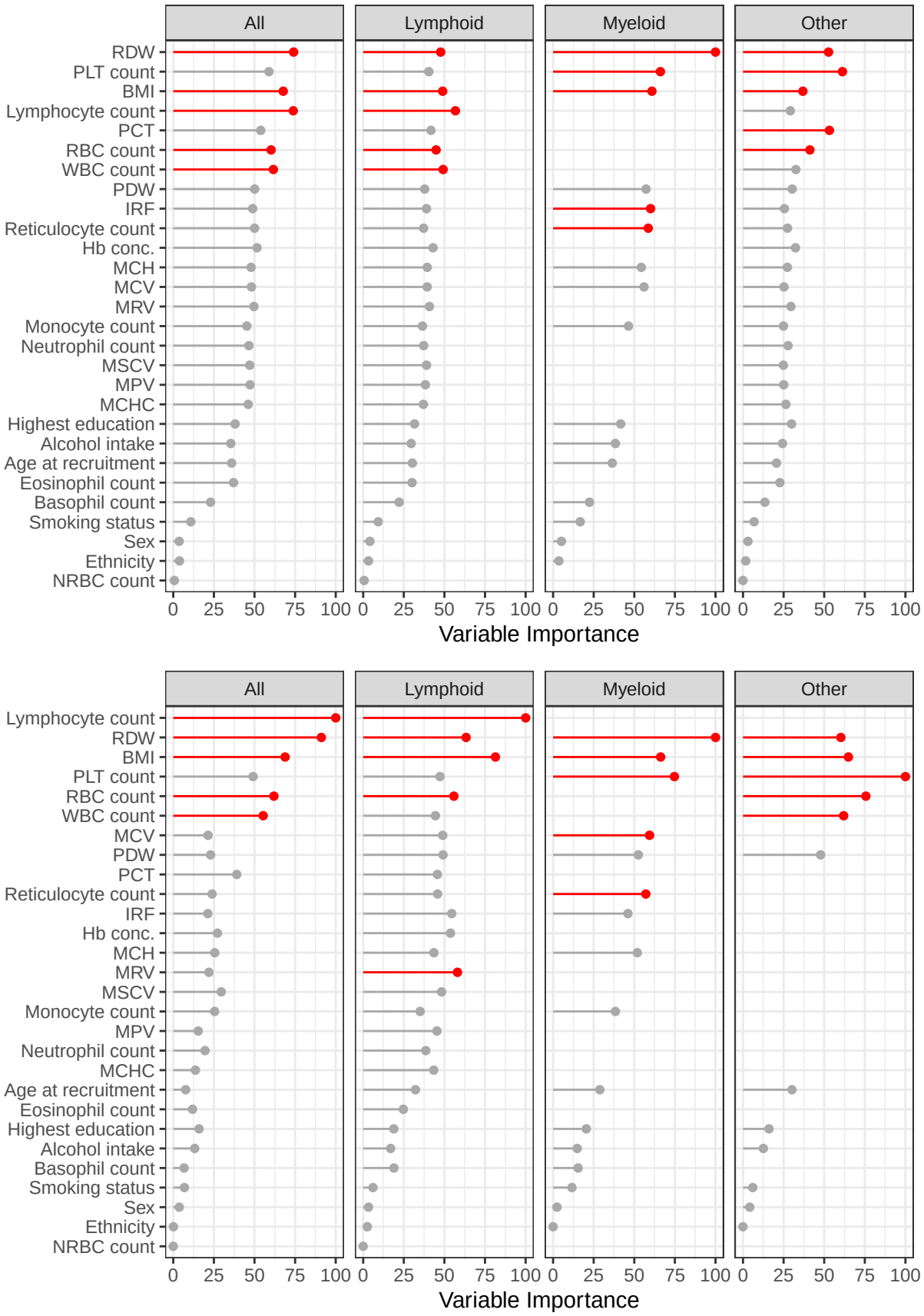


Figure 7. Variable Importance Plots for random forest (top) and XGB (bottom) models. The top 5 variables for all models are coloured in red.

Clustering

The Elbow and Silhouette methods indicate that the optimal number of clusters is 4 (Fig. 7).

The clusters are visualised on a grid using PC1 and PC2 from a PCA of blood counts. Consensus clustering using K-Means identified four distinguishable clusters. Cluster 1 is similar between PAM and consensus K-Means. Cluster 4 from PAM is a combination of cluster 3 and 4 from consensus K-Means. Cluster 2 and 3 from PAM are not clearly separated.

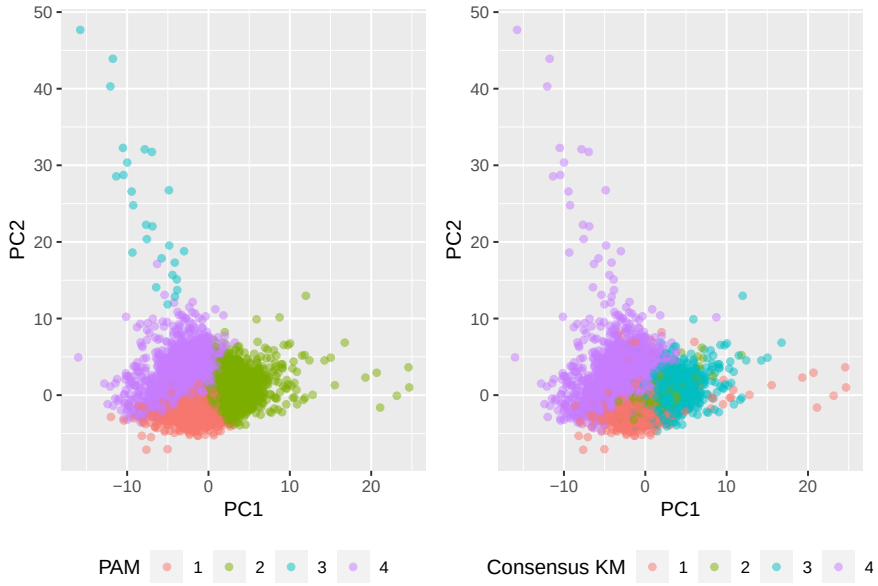


Figure 8. Visualisation of clustering results using PC1 and PC2 from PCA of blood count variables.

Looking at the distribution of blood counts across the clusters (Supplementary Fig. 2 & 3), the variables can be grouped into three large groups, WBC related, RBC related and platelet related. A summary of cluster characteristics for consensus K-Means clusters is presented in Table (x). Cluster 4 is characterised by high WBC related counts while the other clusters have similar levels. Looking at RBC related counts, cluster 1 has large amount of RBC but small size, small variation and low amount of Hb per RBC. Cluster 3 is characterised by small size and low amount of mean Hb but larger size variation. Clusters 2 and 4 have large size and high amount of mean Hb but low size variation. Clusters 2 and 3 have medium levels for all platelet related counts, while cluster 1 tends to have low amount of larger and more variable platelets and cluster 4 has the opposite characteristic, large amount of small and consistently sized platelets.

C-KM	Cluster 1	Cluster 2	Cluster 3	Cluster 4
WBC related count	Low	Low	Low	High
RBC count	High	Low	Medium	Medium
RBC related size	Medium	High	Low	Medium
RBC related size variation	Low	Medium	High	Medium
Hb per RBC	Medium	High	Low	Medium
PLT amount	Low	Medium	Medium	High
PLT size	High	Medium	Medium	Low
PLT size variation	High	Medium	Medium	Low

PAM	Cluster 1	Cluster 2	Cluster 3	Cluster 4
WBC related count	Low	Low	Low	High
RBC count	High	Medium	Low	Medium
RBC related size	Medium	Medium	High	Low
RBC related size variation	Low	Low	Medium	High
Hb per RBC	Medium	High	High	Low
PLT amount	Low	High	Low	High
PLT size	High	Low	High	Low
PLT size variation	High	Low	High	Low

Figure 9. Comparison of blood count levels across clusters.

Summarizing the number of each type of blood cancers in each clusters showed that cluster 4 from both PAM and C-KM has the highest incidence of blood cancer. PAM cluster 2 and C-KM cluster 1 has the lowest incidence.

Table 4. Blood cancer incidence per 1000 participants in each cluster. Some patients are diagnosed with mutliple types of blood cancer.

Method	All	Lymphoid	Myeloid	Other	H_D_cell	N
KM_1	99	61	18	19	1	17825
KM_2	122	72	24	30	1	17068
KM_3	1000	815	0	74	0	27
KM_4	148	75	28	49	2	10926
PAM_1	117	72	23	22	1	14027
PAM_2	79	50	10	19	1	10968
PAM_3	123	73	26	30	1	11673
PAM_4	168	82	33	56	2	9178
C-KM_1	95	59	18	17	1	14844
C-KM_2	119	71	23	30	1	15019
C-KM_3	112	65	21	27	1	8118
C-KM_4	175	89	32	58	2	7865
C-PAM_1	114	70	22	21	1	13370
C-PAM_2	122	72	24	31	1	14899

Method	All	Lymphoid	Myeloid	Other	H_D_cell	N
C-PAM_3	154	77	30	51	2	10777
C-PAM_4	71	46	8	14	0	6800

Discussion

Over all, we demonstrate that tree-based methods have better performance than penalised regression and models using clustering results with RF outperforming XGB. This could be due to the fact that random forest and XGB are ensemble methods which take the consensus from multiple predictors to construct more accurate predictions. Also, non-linear relationships and interactions between variables are taken into account naturally by the tree structures.

The performance of the RF and XGB models determined by AUC value demonstrated that the RF models outperformed XGB models in all categories. Although the classification accuracy for either model is not high enough to compare to human standards and we do not have that information available to compare. The reason for the low accuracy we put forward is that each blood cancer category contains a very broad collection of blood cancer subcategories and performance may be improved a more refined categorization of blood cancers. The XGB boost demonstrates overfitting with the data for sub group and further work needs to be done to optimise the training to prevent this.

Our feature importance plots showed that more features were identified as over the 80% importance in RF models for all cancer, lymphoid and other, but not for myeloid models. This is likely due to the design of the RF and XGB package in that in Random forest has a random selection of features when building trees and so more variables would be part of that selection process whereas in XGB uses boosting to combines weak learners sequentially to allow errors to correct the previous one. This would mean that few variables are selected rather than the random approach used by the RF algorithm.

Clustering models has lower accuracy than models using selected blood counts, the reason could be that by summarising all blood count information into one variable lead to loss of information. Also, the cluster membership may capture information that is unrelated to the outcome investigated.

RDW and PDW are identified by penalised regression to be important feature for all blood cancer types. Both are associated with higher odds of blood cancer, where as the amount of RBC and platelet are associated with lower odds of blood cancer. This suggests that higher variation in RBC and platelet size is associated with a higher risk of blood cancer. These finding as in coherent with previous findings in other cancers(Montagnana and Danese 2016), including breast cancer (Seretis et al. 2013) and colon cancer (Ay et al. 2015).

It is reassuring that similar variables are selected at over 80% in both algorithm and indicate that these are blood markers that should be further assessed and are potential identifiers of future cancer diagnosis however does not indicate if the variable is high or low that could be predicting cancer. The variables selected for lymphoid and myeloid are corresponding to the common cell progenitor are selected as a result. A red cell has a myeloid progenitor and is features highly in both XGB and RF models. The lymphocyte count is selected highly in XGB and RF models in the lymphoid group for the same reason. In the Other group, platelet count is a important feature in prediction which has a myeloid progenitor but the cancer types are heterogenous so it is more challenging to provide a reason why this is selected and should warrant further analysis of these subgroup cancers.

In the RF model All and lymphoid BMI is a important variable to predict blood cancer and is supported with similar findings in the literature that BMI is a risk factor for all cancer (Taghizadeh et al. 2015). It is interesting that although most blood cancers have a genetic cause for them there are lifestyle factors such as BMI as a potential important feature in prediction for blood cancer.

There are no other attempts at predicting broad cancer groups found in the literature to compare our models to although in comparison to human categorisation our models still need significant improvement. Some studies have used blood biomarkers to predict ovarian cancer using a range of ML methods and have similarly found that random forest algorithm performed the best although they have achieved much higher

AUC results than our models likely due to their more narrow objective (Kawakami et al. 2019; Ma et al. 2021).

The best number of clusters was four for all clustering methods, with one group having higher WBC-related counts compared to other groups. The other groups differs in RBC and PLAT count and sizes. The group that has high WBC counts is associated with higher incidence of all types of blood cancer. This suggests that WBC count can reflect a higher risk of blood cancer at least 6 months prior to diagnosis. No groups have been identified to have a differential distribution in blood cancer subtypes, suggesting that the clusters cannot to be used to predict risk of specific types of blood cancer.

Limitations

The difficulty of having a broad base cancer classification approach is that cancer classifications are based on cell lineage affected and the cell type that is affected. We attempted to maintain that classification when we categorized into our 2 groups based on myeloid and lymphoid progenitor cells. It is more difficult with the other group not a homogenous cancer classification group and so is harder to interpret, this does make drawing clear conclusions from findings of Other group more challenging. We would aim to link with genomic data for better understanding of patient specific disease biology and develop precision medical intervention for each patient. This could also be developed into a risk stratification and prognosis but including more data such as treatment modality and deaths for predicting overall survival.

Conclusion

Predictive models were built that can predict blood cancer diagnosis at least 6 months before the event with AUC around 0.7. Random forest has better performance than penalised regression and XGBoost. Clusters were identified based on blood count variables and showed that higher counts of WBCs are related to higher risk of all blood cancer types at least 6 months prior to diagnosis. The results showed that RDW, PDW and increase in WBC counts are important indicators to monitor for blood cancer risk prediction.

Appendix

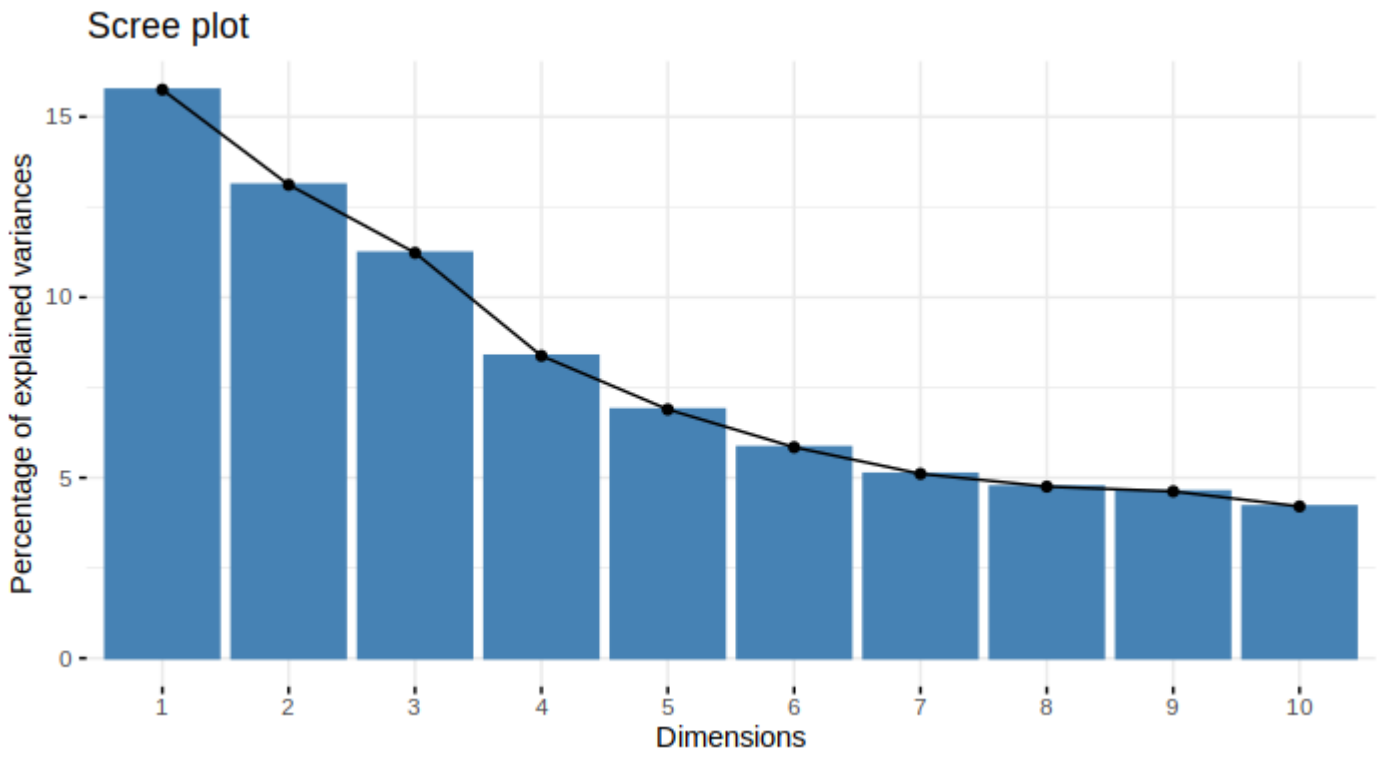
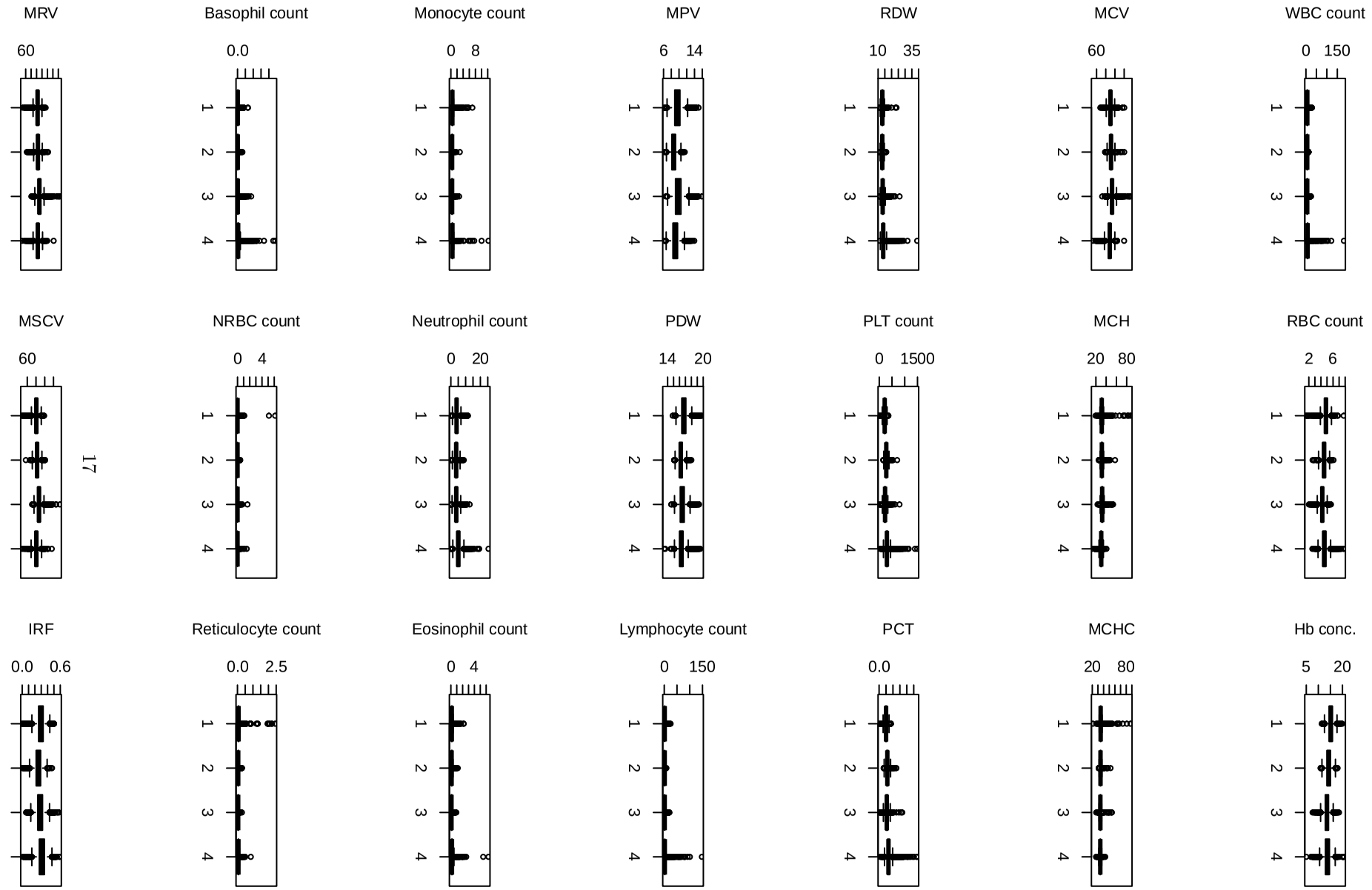
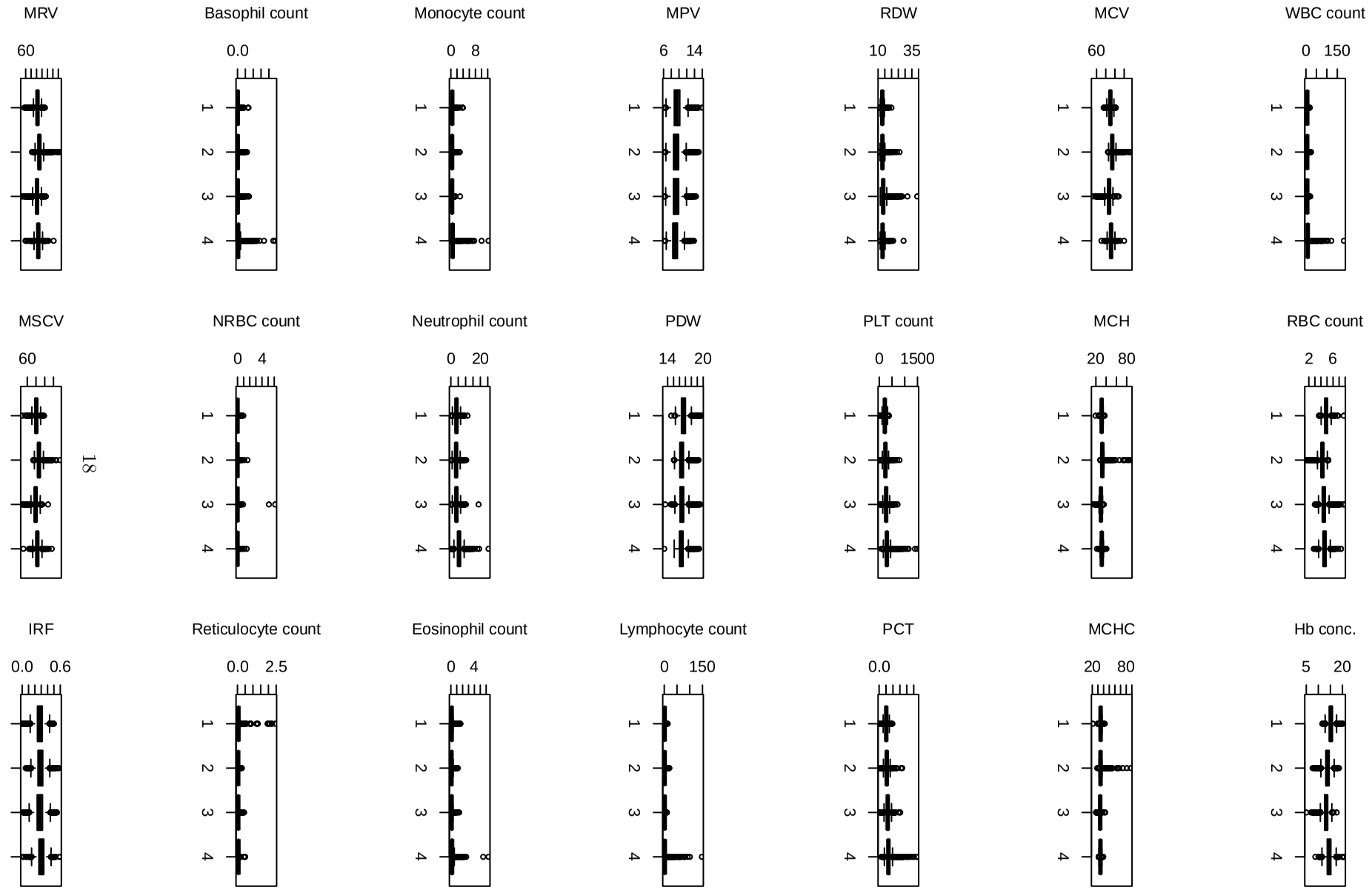


Figure 10. Percentage of variance explained by each PC.


MRV
60
Basophil count
0.0
Monocyte count
0 8
MPV
6 14
RDW
10 35
MCV
60
WBC count
0 150
MSCV
60
NRBC count
0 4
Neutrophil count
0 20
PDW
14 20
PLT count
0 1500
MCH
20 80
RBC count
2 6
IRF
0.0 0.6
Reticulocyte count
0.0 2.5
Eosinophil count
0 4
Lymphocyte count
0 150
PCT
0.0
MCHC
20 80
Hb conc.
5 20
1 2 3 4


MRV
60
Basophil count
0.0
Monocyte count
0 8
MPV
6 14
RDW
10 35
MCV
60
WBC count
0 150
MSCV
60
NRBC count
0 4
Neutrophil count
0 20
PDW
14 20
PLT count
0 1500
MCH
20 80
RBC count
2 6
IRF
0.0 0.6
Reticulocyte count
0.0 2.5
Eosinophil count
0 4
Lymphocyte count
0 150
PCT
0.0
MCHC
20 80
Hb conc.
5 20
1 2 3 4

Reference

Arber, Daniel A., Attilio Orazi, Robert Hasserjian, Jürgen Thiele, Michael J. Borowitz, Michelle M. Le Beau, Clara D. Bloomfield, Mario Cazzola, and James W. Vardiman. 2016. "The 2016 Revision to the World Health Organization Classification of Myeloid Neoplasms and Acute Leukemia." *Blood* 127 (20): 2391–2405. https://doi.org/10.1182/blood-2016-03-643544.

Ay, Serden, Mehmet Ali Eryilmaz, Nergis Aksoy, Ahmet Okus, Yasar Unlu, and Baris Sevinc. 2015. "Is Early Detection of Colon Cancer Possible with Red Blood Cell Distribution Width?" *Asian Pacific Journal of Cancer Prevention* 16 (2): 753–56. https://doi.org/10.7314/APJCP.2015.16.2.753.

Blood Cancer UK. n.d. "Blood Cancer UK | What Is Blood Cancer?" *Blood Cancer UK*. https://bloodcancer.org.uk/understar blood-cancer/what-is-blood-cancer/. Accessed April 10, 2023.

Bodinier, Barbara, Sarah Filippi, Therese Haugdahl Nost, Julien Chiquet, and Marc Chadeau-Hyam. 2021. "Automated Calibration for Stability Selection in Penalised Regression and Graphical Models: A Multi-OMICs Network Application Exploring the Molecular Response to Tobacco Smoking." *arXiv Preprint arXiv:2106.02521*.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Döhner, Hartmut, Andrew H. Wei, and Bob Löwenberg. 2021. "Towards Precision Medicine for AML." *Nature Reviews Clinical Oncology* 18 (9): 577–90. https://doi.org/10.1038/s41571-021-00509-w.

Grinfeld, Jacob, Jyoti Nangalia, E. Joanna Baxter, David C. Wedge, Nicos Angelopoulos, Robert Cantrill, Anna L. Godfrey, et al. 2018. "Classification and Personalized Prognosis in Myeloproliferative Neoplasms." *The New England Journal of Medicine* 379 (15): 1416–30. https://doi.org/10.1056/NEJMoa1716614.

Kawakami, Eiryo, Junya Tabata, Nozomu Yanaihara, Tetsuo Ishikawa, Keita Koseki, Yasushi Iida, Misato Saito, et al. 2019. "Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 25 (10): 3006–15. https://doi.org/10.1158/1078-0432.CCR-18-3378.

Kucine, Nicole. 2020. "Myeloproliferative Neoplasms in Children, Adolescents, and Young Adults." *Current Hematologic Malignancy Reports* 15 (2): 141–48. https://doi.org/10.1007/s11899-020-00571-8.

Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (November): 1–26. https://doi.org/10.18637/jss.v028.i05.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

M. El-Halees, Alaa, and Asem H. Shurrab. 2017. "Blood Tumor Prediction Using Data Mining Techniques." *Health Informatics - An International Journal* 6 (2): 23–30. https://doi.org/10.5121/hiij.2017.6202.

Ma, Jun, Jiani Yang, Yue Jin, Shanshan Cheng, Shan Huang, Nan Zhang, and Yu Wang. 2021. "Artificial Intelligence Based on Blood Biomarkers Including CTCs Predicts Outcomes in Epithelial Ovarian Cancer: A Prospective Study." *OncoTargets and Therapy* 14: 3267–80. https://doi.org/10.2147/OTT.S307546.

Montagnana, Martina, and Elisa Danese. 2016. "Red Cell Distribution Width and Cancer." *Annals of Translational Medicine* 4 (20): 399. https://doi.org/10.21037/atm.2016.10.50.

Seretis, Charalampos, Fotios Seretis, Emmanouil Lagoudianakis, George Gemenetzis, and Nikolaos S. Salemis. 2013. "Is Red Cell Distribution Width a Novel Biomarker of Breast Cancer Activity? Data From a Pilot Study." *Journal of Clinical Medicine Research* 5 (2): 121–26. https://doi.org/10.4021/jocmr1214w.

Shahid, Afzal Hussain, and M. P. Singh. 2019. "Computational Intelligence Techniques for Medical Diagnosis and Prognosis: Problems and Current Developments." *Biocybernetics and Biomedical Engineering* 39 (3): 638–72. https://doi.org/10.1016/j.bbe.2019.05.010.

Taghizadeh, Niloofar, H. Marike Boezen, Jan P. Schouten, Carolien P. Schröder, E. G. Elisabeth de Vries, and Judith M. Vonk. 2015. "BMI and Lifetime Changes in BMI and Cancer Mortality Risk." *PloS One* 10 (4): e0125261. https://doi.org/10.1371/journal.pone.0125261.

Wang, Qiwei, Shusheng Bi, Minglei Sun, Yuliang Wang, Di Wang, and Shaobao Yang. 2019. "Deep Learning Approach to Peripheral Leukocyte Recognition." *PloS One* 14 (6): e0218808. https://doi.org/10.1371/

journal.pone.0218808.

Yang, Jin Hyuk, Yonggoo Kim, Jihyang Lim, Myungshin Kim, Eun-Jee Oh, Hae-Kyung Lee, Yeon-Joon Park, et al. 2014. "Determination of Acute Leukemia Lineage with New Morphologic Parameters Available in the Complete Blood Cell Count." *Annals of Clinical & Laboratory Science* 44 (1): 19–26.

Yao, Hongxia, Liyou Lian, Rujie Zheng, and Chen Chen. 2023. "Red Blood Cell Distribution Width/Platelet Ratio on Admission as a Predictor for in-Hospital Mortality in Patients with Acute Myocardial Infarction: A Retrospective Analysis from MIMIC-IV Database." *BMC Anesthesiology* 23 (1): 113. https://doi.org/10.1186/s12871-023-02071-7.