IMPROVED BIOMARKER-BASED DIAGNOSTICS OF LEUKEMIA SUBTYPES USING MACHINE LEARNING METHODS

Katherine Schmidt[1], Purnima M. Kodate[2,] and Kirti M. Yenkie[3*]
[1]Department of Mathematics, Rowan University, Glassboro, NJ, USA
[2]Department of Pathology, Government Medical College, Nagpur, India
[3]Department of Chemical Engineering, Rowan University, Glassboro, NJ, USA

Abstract

With cancer being a leading cause of death worldwide, it is of utmost importance to detect the disease at an early stage and implement suitable treatment strategies to avoid the disease progression as well as minimize any other health complications resulting from incorrect diagnosis and inefficient treatment. One possible way to detect cancer early and its correct sub-type is through the use of biomarkers. This work focuses on the classification of leukemia, specifically Acute Myeloid Leukemia (AML), using machine learning. The two algorithms that are explored in this work include decision trees and artificial neural networks. Myeloperoxidase (MPO) is the main biomarker of focus in the classification and prognosis of AML. The two machine learning methods showed that MPO was a significant factor in the classification of AML in our dataset, as the accuracies drop significantly when this variable is removed from the model. Other variables studied include age, gender, periodic acid-Schiff, and total leukocyte count.

Keywords

Acute myeloid leukemia, myeloperoxidase, machine learning, computational biology, diagnostics

Introduction

Cancer is a leading cause of death worldwide. Approximately 1 in 285 US children will be diagnosed with cancer before their twentieth birthday (American Cancer Society, Surveillance Research, 2014). Leukemia, particularly ALL (Acute Lymphoblastic Leukemia) and AML (Acute Myeloid Leukemia), are the most common type of childhood cancers. Leukemias and blood cancers are not only limited to children; ALL and AML are also common in adults, especially in the population over fifty-years of age (Niederhuber et al., 2013).

Early diagnosis is extremely important in designing targeted therapies with minimal side-effects. Identification of key biomarkers which are indicators of the disease state can provide insights into the molecular and cellular mechanisms which drive cancer proliferation and any associated effects (Kalia, 2015). Biomarkers can be studied at various levels including genomics, transcriptomics, and proteomics levels to provide better diagnosis and treatment for several cancers, particularly leukemia subtypes such as ALL, AML, CLL (Chronic Lymphocytic Leukemia), etc (Mirzaei et al., 2018). Some of them can also provide insights into potential co-morbidities, especially cardiac stress and cardiovascular diseases occurring along with cancer or due to its therapeutic side-effects (Lipshultz et al., 2012; Mair et al., 2015). Myeloperoxidase is one such biomarker that can be used clinically for efficient diagnosis and quantification of

* To whom all correspondence should be addressed
Email: yenkie@rowan.edu (K. M. Yenkie)

associated cardiovascular risks in cancer patients (Knapp et al., 1994; Loria et al., 2008).

MPO has not been included in many previous studies. However, studies did find a link between MPO and patient's six-month outcome of acute coronary syndromes (Baldus, 2003) as well as identifying a relationship between MPO levels in patients with AML and whether or not they would benefit from a transplant (Kim et al., 2012). This means not only can MPO provide information as a diagnosis factor, but it may be useful in prognosis as well.

To this end, two machine learning algorithms; decision trees, and artificial neural networks were applied to examine the relationship between MPO as a biomarker of AML. The models were fit at numerous training and testing sets as well as tested multiple scenarios to include the overall best fit model.

Data

The diagnostic data utilized in this work is provided by our clinical collaborator, Dr. P. M. Kodate (MD Pathology) from the Government Medical College in Nagpur, India. The data consists of 174 observations of 15 variables. Of these variables, ten are binary, four numerical, and dependent variables (immunophenotyping diagnosis) is categorical with five levels, namely AML, ALL, mixed-phenotype acute leukemia (MPAL), normal, and inconclusive. The independent variables include age, gender, hepatomegaly, splenomegaly, mediastinal mass, bleeding, total leucocyte count, hemoglobin levels, platelet count, periodic acid Schiff, and MPO.

Of all the patients studied, 77 have an immunophenotyping diagnosis of AML, 92 are diagnosed as ALL, two inconclusive, two normal, and one MPAL. One other significant factor to consider is the cutoff for "positive" MPO. MPO is noted as positive if more than 3% of MPO-positive blasts are present out of all the leukocytes.

Methods

A decision tree and neural network were built in order to identify significant factors and examine the role of MPO in identifying the immunophenotyping diagnosis of AML. Both the decision tree and artificial neural network (ANN) were fit to four different training and testing set splits ranging from 65% to 85% training data and the remaining testing. Furthermore, each algorithm was fit twice: once including MPO and once removing MPO as a variable to determine its overall significance in the accuracy of the model and leukemia subtype prediction.

Decision Trees

A total of eight decision trees were fit at four different training and testing dataset splits as well as with and without MPO as a variable in the model. The Gini coefficient was used as splitting criteria, where a lower Gini indicated a higher gain, signifying the lowest Gini would determine the best fit. In order to avoid overfitting the model to the training data, each tree was pruned using a k-fold cross validation (Blockeel and Struyf, 2002) with a k of ten.

Artificial Neural Networks

Just as with the decision trees, eight artificial neural networks were fitted. There were two cases, either with or without MPO and four different training and testing sets, resulting in a total of eight models. The first step to creating these models was splitting the data into the training and test sets. Then, contrary to the decision tree, the numeric variables had to be normalized. This means the variables were put on the same scale, thus were rescaled to have values between zero and one. This is important to ensure that no variable is being weighted higher than another due to its numeric value rather than its significance in the model.

The next step was fitting the model. The model was tested at four different training-testing splits as well as with one, two, or three hidden layers in each. This was done to achieve the most accurate model possible. After each model was fitted, the accuracy was calculated using the testing data. The model with the highest testing accuracy was deemed to be the most optimal model. The most accurate model resulted in only one hidden node, allowing for easy-interpretability of the weights each variable holds in the model.

Conclusions

Both the algorithms resulted in over 90% accuracy for predicting AML. The resulting accuracies of the best fit models both with and without MPO for each method are summarized in Table 1 below.

Table 1: Accuracies of the best fit models

	Decision Tree	**ANN**
With MPO	97.14%	100%
Without MPO	91.43%	90.91%

It is clear that in both cases, the model with MPO yielded significantly better results than the model without MPO. It is verified with our models that the diagnosis of AML is most affected by the presence or absence of myeloperoxidase. With MPO included in the models, the decision tree and neural network performed at 97.14% and 100% accuracy, respectively. However, when MPO is removed from the model, those accuracies drop to 91.43% and 90.91%. Within the dataset itself, 78% of AML

patients are noted with the presence of MPO, while no other diagnosis contains any patients with a presence of MPO. This suggests that MPO is a significant indicator in detecting AML, but this may not be the case for ALL, which makes up approximately half of the data points. Another possibility is the result of increased MPO levels due to cardiac damage from the chemotherapy treatment in patients with AML. In order to take this observation into consideration, it would be important to know the history of treatment of the patients and current conditions of other cardiac biomarkers.

MPO may also be useful as a prognosis factor for AML. A study by the Eastern Cooperative Oncology Group (ECOG) found that patients with more than 50% MPO-positive blasts showed a higher complete response (CR) rate than those with less than 50% MPO-positive blasts (Matsuo et al., 1989). The National Cancer Institute (NCI) defines CR as the disappearance of all signs of cancer in response to treatment (NCI Dictionary of Cancer Terms, 2011). This coincides with the aforementioned study (Kim et al., 2012), which suggest MPO may be valuable in identifying those who will benefit from a stem cell transplant. These studies focus on MPO as a prognosis factor, which may be an important application of the biomarker outside of classification.

Future studies may be able to incorporate a larger number of biomarkers in their analysis, such as troponin proteins, myeloperoxidase, and natriuretic peptides. It is rare that a study includes all of these in the analysis. However, these measurements can be taken in a relatively easy manner and may lead to great insights as not only important diagnosis factors but also aid in optimal prognosis.

Acknowledgements

The authors thank the Department of Pathology at the Government Medical College and Hospital, Nagpur, India for providing the resources to collect clinical data and the Rowan University Institutional Review Board (IRB) for approving the study protocol.

References

American Cancer Society, Surveillance Research, 2014. American Cancer Society (ACS) Special section: Cancer in children and Adolescents.

Baldus, S., 2003. Myeloperoxidase Serum Levels Predict Risk in Patients With Acute Coronary Syndromes. Circulation 108, 1440–1445. https://doi.org/10.1161/01.CIR.0000090690.67322.51

Blockeel, H., Struyf, J., 2002. Efficient Algorithms for Decision Tree Cross-validation 30.

Kalia, M., 2015. Biomarkers for personalized oncology: recent advances and future challenges. Metabolism 64, S16-21. https://doi.org/10.1016/j.metabol.2014.10.027

Kim, Y., Yoon, S., Kim, S.J., Kim, J.S., Cheong, J.-W., Min, Y.H., 2012. Myeloperoxidase Expression in Acute Myeloid Leukemia Helps Identifying Patients to Benefit from Transplant. Yonsei Med J 53, 530–536. https://doi.org/10.3349/ymj.2012.53.3.530

Knapp, W., Strobl, H., Majdic, O., 1994. Flow cytometric analysis of cell-surface and intracelluar antigens in leukemia diagnosis. Cytometry 18, 187–198. https://doi.org/10.1002/cyto.990180402

Lipshultz, S.E., Miller, T.L., Scully, R.E., Lipsitz, S.R., Rifai, N., Silverman, L.B., Colan, S.D., Neuberg, D.S., Dahlberg, S.E., Henkel, J.M., Asselin, B.L., Athale, U.H., Clavell, L.A., Laverdière, C., Michon, B., Schorin, M.A., Sallan, S.E., 2012. Changes in Cardiac Biomarkers During Doxorubicin Treatment of Pediatric Patients With High-Risk Acute Lymphoblastic Leukemia: Associations With Long-Term Echocardiographic Outcomes. JCO 30, 1042–1049. https://doi.org/10.1200/JCO.2010.30.3404

Loria, V., Dato, I., Graziani, F., Biasucci, L.M., 2008. Myeloperoxidase: A New Biomarker of Inflammation in Ischemic Heart Disease and Acute Coronary Syndromes. Mediators of Inflammation 2008. https://doi.org/10.1155/2008/135625

Mair, J., Jaffe, A., Apple, F., Lindahl, B., 2015. Cardiac Biomarkers. Dis Markers 2015. https://doi.org/10.1155/2015/370569

Matsuo, T., Cox, C., Bennett, J.M., 1989. Prognostic significance of myeloperoxidase positivity of blast cells in acute myeloblastic leukemia without maturation (FAB: M1): an ECOG study. Hematol Pathol 3, 153–158.

Mirzaei, H., Fathullahzadeh, S., Khanmohammadi, R., Darijani, M., Momeni, F., Masoudifar, A., Goodarzi, M., Mardanshah, O., Stenvang, J., Jaafari, M.R., Mirzaei, H.R., 2018. State of the art in microRNA as diagnostic and therapeutic biomarkers in chronic lymphocytic leukemia. Journal of Cellular Physiology 233, 888–900. https://doi.org/10.1002/jcp.25799

NCI Dictionary of Cancer Terms [WWW Document], 2011. . National Cancer Institute. URL https://www.cancer.gov/publications/dictionaries/cancer-terms (accessed 3.12.19).

Niederhuber, J.E., Armitage, J.O., Doroshow, J.H., Kastan, M.B., Tepper, J.E., 2013. Abeloff's Clinical Oncology: Fifth Edition. Elsevier Inc.