Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems

Enrique J. deAndrés-Galiana[1,4,*], Juan L. Fernández-Martínez[1,*,§], Oscar Luaces[4], Juan J. del Coz[4], Leticia Huergo-Zapico[3], Andrea Acebes-Huerta[3], Segundo González[3], Ana P. González-Rodríguez[2]

[1]Department of Mathematics, University of Oviedo, Spain.
[2]Hematology Department, Hospital Central de Asturias, Oviedo, Spain.
[3]Instituto Universitario Oncológico del Principado de Asturias (IUOPA). University of Oviedo, Spain.
[4]Artificial Intelligence Center, University of Oviedo, Spain.
* Both authors have contributed equally to this study.
[§]Corresponding author

Corresponding author: JLFM; jlfm@uniovi.es; address: Jesús Arias de Velasco s/n 33005 Oviedo, Spain, 0034 985 103 199.

Abstract

Introduction: Chronic Lymphocytic Leukemia (CLL) is a disease with highly heterogeneous clinical course. A key goal is the prediction of patients with high risk of disease progression, which could benefit from an earlier or more intense treatment. In this work we introduce a simple methodology based on machine learning methods to help physicians in their decision making in different problems related to CLL. **Material and Methods**: Clinical data belongs to a retrospective study of a cohort of 265 Caucasians who were diagnosed with CLL between 1997 and 2007 in Hospital Cabueñes (Asturias, Spain). Different machine learning methods were applied to find the shortest list of most discriminatory prognostic variables to predict the need of Chemotherapy Treatment and the development of an Autoimmune Disease. **Results**: Autoimmune disease occurrence was predicted with very high accuracy (>90%). Autoimmune disease development is currently an unpredictable severe complication of CLL. Chemotherapy Treatment has been predicted with a lower accuracy (80%). Risk analysis showed that the number of false positives and false negatives are well balanced. **Conclusions**: Our study highlights the importance of prognostic variables associated with the characteristics of platelets, reticulocytes and natural killers, which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia for autoimmune disease development, and also, the relevance of some clinical variables related with the immune characteristics of CLL patients that are not taking into account by current prognostic markers for predicting the need of chemotherapy. Because of its simplicity, this methodology could be implemented in spreadsheets.

Keywords: Chronic Lymphocytic Leukemia; chemotherapy treatment; autoimmune disease development; machine learning.

1. INTRODUCTION

Chronic Lymphocytic Leukemia (CLL) is the most common adult Leukemia in western countries, and it is characterized by the accumulation of malignant B-cells in blood and lymphoid organs. The clinical course of CLL is highly heterogeneous since the survival of some patients is only slightly affected by the disease, whereas other patients have a progressive disease associated with infectious and autoimmune complications. These progressive patients have poor prognosis, but they could benefit from an earlier or more intense chemotherapeutic treatment. It has been reported that many poor prognostic factors (including CD38, ZAP-70, β2-microglobulin, IgVH mutation status and deletions of 11q23 or 17p53) may help to identify high-risk patients at early stages [1 - 6]. Most of these prognostic factors focus on the analysis of the characteristics of malignant leukemia cells. Additionally, the characteristics of the immune system of CLL patients, such as the number of CD8 and CD4-T cells at diagnosis, may also predict the progression of the disease [6]. Nevertheless, due to their high cost and complexity some of these prognostic factors are not used in most hospitals on regular basis. To overcome this problem in the clinical practice staging systems using few, simple, cheap and accessible clinical variables have been popularized. The Rai staging system [7] and the Binet classification [8] are useful to predict the prognosis of CLL patients, to stratify them, and to achieve comparisons for interpreting specific treatment results. Staging systems stratify subsets of patients who have significant differences in the overall survival but they fail to identify patients who have a high risk of progression in early stages of the disease. Additionally, no current prognostic factors exist to predict the development of some severe complications such as the development of Autoimmune Diseases (AD), or the need for chemotherapy. Consequently, the identification of currently available clinical variables to assess the medical decisions in these CLL-related diagnosis problems is a key goal in the management of this disease. The development of AD or the need of CT is not known at diagnosis. So far, only with the evolution of the patient during the 5 years follow up, medical doctors can answer these questions. Therefore, the interest of the methodology presented herein consists in being able of predicting both CLL related problems at diagnosis. Particularly, AD problem was very hard to predict, and up to our knowledge no previous research was successful to explain this phenomenon using biochemical variables.

In this paper we show whether machine learning methods and clinical data

obtained from a large population of well-studied CLL patients [6] can be efficiently applied to address these CLL diagnosis problems in medical practice by capturing the hidden implicit relationships between the clinical variables and the corresponding class of the different patients that have been established by medical experts. The use of machine learning techniques [9] in clinical medicine [10] and in cancer prediction and prognosis [11] is not new, and it has the advantage of treating more general prediction problems than survival analysis (usually treated through the Kaplan-Meier estimator) as supervised classification problems that admit more stable solutions than the corresponding regression problems.

The machine learning methodologies that are proposed in this paper are simple in their design and serve to provide to the physicians a simple and robust decision-making support system. Other more complex algorithms could be used, but the goal of this work is to obtain a simple decision rule and not to compare different learning algorithms. This manuscript is structured in three main parts. Firstly we provide an exhaustive explanation of the methods. Secondly, we present the results obtained for the two clinical CLL-related problems addressed herein: need of Chemotherapy Treatment (CT) and Autoimmune Disease development (AD). Finally, we provide coherent explanations and discussion of the findings.

2. MATERIAL AND METHODS

A cohort of two hundred sixty-five Caucasians who were diagnosed in the Cabueñes Hospital (Gijón, Spain) with CLL between 1997 and 2009 were enrolled in this study. The population distribution by gender and age is the following: 154 are males and 111 are females, with ages ranging from 42 to 92, and 47 to 94 years old respectively. Clinical characteristics of patients including time for diagnosis to first treatment, need of chemotherapy treatment and appearance of autoimmune complications were also taken into account in this study. Additionally, thirty-six different clinical and biological variables were measured at diagnosis of the disease. Table 1 shows the variables description used in this study. Some variables reflect the malignant characteristic of leukemia cells; others measure the immunological characteristics of CLL patients, and some may be associated with the presence or development of autoimmune complications (autoimmune haemolytic anemia and immune-thrombocytopenia). Finally, some of the variables are demographic and biochemical. Most of them have a sampling frequency higher than 80%, however, the

reticulocyte count (RET) and ZAP-70 are the ones that show the lowest sampling frequency. Particularly, ZAP-70 is only sampled in 21.9% of the patients (58 out of 265), showing that this popular CLL prognostic factor is not always available in medical practice. Although some of these variables were not at disposal at diagnosis (LD for instance), they have been used for analysis purposes. We provide the database as supplementary material (see "CLL.xls").

The problems to be solved in this manuscript are the prediction of the need for Chemotherapy Treatment (CT) and the development of Autoimmune Disease (AD). Both classification problems are binary (two class classification problem). In our methodology we have explored the minimum-size list of prognostic variables (named as reduced base) having the highest predictive accuracy using different feature selection methods. The selected prognostic variables will be subsequently used for diagnosis and prognosis.

Figure 1 shows the flowchart of the methodology, that includes 4 different steps:

2.1. Data Preprocessing

Data preprocessing is applied to improve the quality of data used for performing feature selection, prediction and optimization. It includes two main sub steps that can be applied or not depending on their impact on the prediction:

- *Filtering*: All the features that were sampled less than a certain sampling frequency are removed. The filtering cut offs used were 30, 40 and 50%.
- *Imputation*: This technique consists in interpolating all the missing values using a Nearest-Neighbor algorithm [12]. Given a partially-informed sample (with missing values) the algorithm finds the closest sample within the set of fully-informed samples and gives the values of the missing variables in this closest sample to the imputed sample. The similarity between samples is measured using the standard Euclidean dot product in N-dimensional vector spaces, where N is the number of fully-informed variables. This way of interpolation has the advantage of not introducing additional outliers that are not originally present in the dataset before imputation. Although the success of the different imputed algorithms might be data-driven, imputing the data improved the accuracy in the predictions and did not alter the prognostic variables that were involved providing shorter lists with higher discriminatory power.

2.2. Feature Selection methods

Maximum Fisher's ratio [13,14]: The Fisher's ratio of an attribute *j*, in a two-class problem, c_1, c_2, is defined as follows:

$$GFR_j(c_1,c_2) = \frac{\left(\mu_{j1} - \mu_{j2}\right)^2}{\sigma_{j1}^2 + \sigma_{j2}^2},$$

where, μ_{j1}, μ_{j2} are measures of the center of the distribution (means) of gene j in classes 1 and 2, and $\sigma_{j1}^2, \sigma_{j2}^2$ are measures of the dispersion (variance) within these classes. This method looks for prognostic variables that separate the classes further apart and are very homogeneous within classes (low intra class variance).

Minimum class Entropy [15, 16]: Entropy is a measure of the number of specific ways in which a system may be rearranged, and it is often considered a measure of disorder, or progression towards thermodynamic equilibrium. In the case of a binary classification problem, the entropy of each attribute is defined as follows:

$$E_j(c_1,c_2) = -\sum_{k=1}^{2}\sum_{j=1}^{N_C} p_{kj} \log_2 p_{kj},$$

where N_C are the number of bins used to describe the probability distribution of attribute *j* in class *k,* and p_{kj} is the probability that this attribute takes the center class value x_{kj}. The algorithm to compute the entropy is based in ordering the variables according to their value and calculating the mismatch to the class vector. A perfect ordering occurs when the values correspond perfectly to the class vector. Variables with higher ordering (or lower entropy) are therefore the most discriminatory.

Maximum Percentile Distance: This feature selection method selects the attributes with higher distances between the corresponding cumulative probability functions (percentile array) within each class, defined for attribute *j* as follows:

$$d_j(c_1,c_2) = \frac{\left\|\mathbf{p}_{j1} - \mathbf{p}_{j2}\right\|_2}{\max\left(\left\|\mathbf{p}_{j1}\right\|_2, \left\|\mathbf{p}_{j2}\right\|_2\right)},$$

where $\mathbf{p}_{ji}$ stands for the percentile vector *j* in class *i*, and $\left\|\mathbf{p}_{ji}\right\|_2$ its Euclidean

norm. Percentiles vary from 5 to 95 to avoid the possible effect of outliers [17]. This method can be considered as a generalization of a Mann-Whitney selection test, which is only based in the median (percentile 50).

The main reason for choosing these methods is due their clear interpretation, low computational cost, and the possibility of being applied to both, discrete and continuous variables. A survey about FS methods can be consulted in [18].

2.3. Accuracy evaluation

Once the most discriminatory variables are determined and ranked in decreasing order by their discriminatory power, the aim is to determine the shortest (having the smallest number of variables) list of prognostic variables with the highest predictive accuracy. The algorithm to find the minimum-size list of features is the Backwards Feature Elimination (BFE), which is similar to the Recursive Feature Elimination [19]. Feature elimination tries to unravel the existence of redundant or irrelevant features to yield the smallest set of prognostic variables that provide the greatest possible classification accuracy. Redundant features are those that provide no additional information than the currently selected features, while irrelevant features provide no useful information in any context.

The algorithm of BFE works as follows:

1. Beginning by the tail of the ranked list of prognostic variables, the algorithm iteratively generates increasingly shorter lists by eliminating one prognostic variable at a time, calculating their classification accuracy.
2. Finally, the list with the optimum accuracy and minimum size is therefore selected.

This way of proceeding is based on the following idea: prognostic variables with higher discriminatory ratios span low frequency features of the classification, whilst variables with lowest discriminatory ratios account for the details in the discrimination (high frequency features). This method determines the minimum amount of high frequency details that are needed to optimally discriminate between classes.

The predictive accuracy estimation is based on a Leave One Out Cross-Validation experiment (LOOCV), using the average distance of the reduced set of features to each training class set. The goal of cross-validation is to estimate how accurately a predictive model (classifier) will perform in practice. LOOCV involves using a single sample from the original dataset as the validation data (sample

test), and the remaining samples as training data. The class assignment is based in a nearest-neighbor classifier in the reduced base, that is, the class with the minimum distance in the reduced base to the sample test is assigned to the sample test. The average LOOCV predictive accuracy is calculated by iterating over all the samples using as metric the Euclidean distance between the corresponding normalized variables. For that purpose the weights used to normalize the variables are the inverse of two times the prior variability (standard deviation) of the prognostic variables. These weights serve to scale the different kinds of measurements into approximately the same range in order to give to each variable a similar influence on the overall distance measurement. The distance between a new sample $\mathbf{s}_{new}$ and the average signature $\mathbf{m}_j$ in class *j* is:

$$d\left(\mathbf{s}_{new}, \mathbf{m}_j\right) = \left\| W\left(\mathbf{s}_{new} - \mathbf{m}_j\right)\right\|_2,$$

with W is a diagonal matrix with $W(k,k) = \dfrac{1}{2std(v_k)}$, where $std(v_k)$ is the standard deviation of the *k-th* discriminatory prognostic variable.

In this procedure the feature selection method is executed only once using all training samples before estimating the accuracy by means of a leave-one-out procedure. For each new sample the classifier computes the average distance to the training samples of each class, being d_1 the average distance to class 1, and d_2 the average distance to class 2.

Based on these distances the probability of a new sample $\mathbf{s}_{new}$ to be in class 1 can be written as:

$$P\left(\mathbf{s}_{new} \in c_1\right) = \frac{d_2}{d_1 + d_2}.$$

The procedure to decide the class assignment is as follows:

$$\mathbf{s}_{new} \in c_1 \text{ if } P\left(\mathbf{s}_{new} \in c_1\right) > p_{th} = 0.5.$$

Otherwise, $\mathbf{s}_{new} \in c_2$. The threshold probability $\left(p_{th}\right)$ can be considered as a continuous variable to establish the Receiver Operator Characteristic (ROC) curve for this classifier [20]. Finally, the reduced base might be tested over different randomly chosen training and testing dataset, and averaging the results over a set of independent simulations.

Although this simple classifier seems to be similar to a nearest neighbor algorithm (k-NN), it is not obviously the same, since neither the centroid definition of the distributions, nor the way of adopting the decisions coincide. Besides, we have testing k-NN nearest neighbor classifiers without success. Notice that in this process, the feature selection method is executed only once using all training samples, before estimating the accuracy by means of a leave-one-out procedure. Our goal is to study the effectiveness of feature selection methods in finding the groups of prognosis variables with higher predictive accuracy of these two CLL-related problems. Also, if the attribute selection process was performed each time the classifier was executed (i.e. in each of the folds of the leave-one-out), different sets of attributes would be obtained, thus, it would more difficult to assess the goodness of any concrete group of prognosis variables. The only way will be performing frequency analysis of the selected prognostic variables and applying BFE to this set of variables ranked by decreasing order of their posterior frequency. Besides, since the accuracy is established by Leave-One-Out Cross Validation (LOOCV) the selected attributes within each fold of the LOOCV would not be so different from selecting them using the whole dataset, considering that the training set of each of fold in a LOOCV is composed by all the samples but one. These facts have been confirmed through numerical experimentation.

2.4. ROC curves and risk assessment

In the previous step, maximizing the predictive accuracy according to the LOOCV criterion allowed to determine the best reduced-base of prognostic variables. However, it is also important to analyze the structure of the confusion matrix, obtained from the set of predictions of the training set using the LOOCV method. The confusion matrix is composed by: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These concepts depend on how the classification problem is set up. From the confusion matrix we can calculate different rates that are very useful to understand the risk in the prediction:

- True Positive Rate or Sensitivity (TPR): measures the proportion of actual positives that are correctly predicted as such.
- True Negative Rate or Specificity (SPC): measures the proportion of negatives that are correctly predicted as such.

- Positive Predicted Value (PPV): is the proportion of positives values that are true positives.
- False Positive Rate (FPR): fraction of false positives out of the total actual negatives.
- False Negative Rate (FNR): fraction of false negatives out of the total actual positives.
- False Discovery Rate (FDR): fraction of false positives out of the total actual positives.

Based in these rates it is possible to construct a receiver operating characteristic curve (or ROC curve), which is a graphical plot that illustrates the performance of a binary classifier as a function of one parameter (the cut-off probability in this case). The curve is created by plotting the true positive rate or sensitivity (TPR) against the false positive rate (FPR) or fall-out. A perfect classifier has as ROC curve the step function at the origin. ROC analysis is related to cost/benefit analysis of diagnostic decision making (see for instance [17]).

The selected attributes are used to provide simple biomedical discriminatory rules for diagnosis and prognosis since for each classification problem we provide the bounds for the four groups of the confusion matrix. This knowledge can be used by the physicians in their decision-making process. Additionally to the LOOCV results, we also provide the mean accuracy obtained for 100 random holdouts 75/25 (75% for training and 25% for testing). In any case, and independently of how the predictive accuracy is established, it is crucially important to understand that there exist different combinations of prognostic variables with similar predictive accuracy whose knowledge might be useful to understand the genesis of the problem from a medical point of view. The existence of these different lists is related to the uncertainty analysis of the solutions in any decision-making problem [21 -22].

Finally, the aim of this paper is not to compare different machine learning methods, but to introduce a simple methodology to select the shortest list of prognostic variables that could be easily interpreted by medical doctors to perform prognostic predictions with their corresponding risk assessment. However, we have compared this distance based nearest-neighbor algorithm to more sophisticated learning methods and the results did not improve or were clearly worse. The success of the methodology is not based on the sophistication of the classifier but on selecting

the most discriminatory variables in each case and building the classifier based on these variables. By doing that it has been shown that the classification problem becomes linearly separable [23].

The methodology presented herein is easy to understand, since we avoid the use of black-box methodologies that provide estimations without MD's understanding, and has been successfully applied to predict response to treatment in Hodgkin lymphoma [17] using clinical data, and also in the prediction of risk of radiotherapy-related fatigue in prostate cancer patients using high dimensional expression data [24].

3. RESULTS AND DISCUSSION

3.1. Chemotherapy Treatment Assessment

As it was already mentioned CLL has a highly variable clinical course. Some patients have an indolent disease and they do not require CT. Other patients who present a progressive disease may require an intense CT. The identification of those patients at early stages of the disease with a high risk of rapid disease progression may help to significantly improve their prognosis. Thus, we try to establish the prognostic variables and criteria to assess the need for CT, assuming that the clinical decisions on the 71 (out of 259, therefore there are 6 missing values since the total cohort is 265) patients that have received CT were correct. The criteria for initiating CT were established in 2008 by the International Workshop on Chronic Lymphocytic Leukemia [25]. Particularly the presence of constitutional symptoms, such as, unintentional weight loss of 10% or more within the previous 6 month and significant fatigue or fevers or night sweats without other evidence of infection.

The Fisher's ratio method provided the minimum-size set of prognostic variables with the highest accuracy of 80.3%: B2M, WBC, ALC and MBC. Figure 2 shows the ROC curve and the Recall (or True Positive Rate -TPR) against Precision (or Positive Predicted Value - PPV) curves for several probability thresholds in the CT classification problem. The optimum result ($p_{th} = 0.47$) shows that 63.4% (TPR) of the patients that need CT and 86.7% (True Negative Rate or Specificity – SPC) of the patients that do not need CT were correctly predicted. Besides, with that probability threshold we got a Precision (or Positive Predicted Value – PPV) of 64.3%. Nevertheless, other probability thresholds could be adopted depending on the Recall/Specificity balance, and therefore on the PPV as well. The False Discovery

Rate (FDR) was 36.62%. The confusion matrix is shown below:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 45 & 25 \\ 26 & 163 \end{pmatrix}.$$

The True Positives (TP) are formed by the group of patients that need CT (+) and are correctly predicted, and the True Negatives (TN) are formed by the groups of patients that do not need CT (-) and are correctly predicted. Thus, False Positives (FP) are the patients that do not need CT (-) and are not correctly predicted and False Negative (FN) are the patients that need CT (+) and are not correctly predicted.

Besides, we have performed a two sample T-test and a Mann-Whitney U-test to clarify the differences between the TP and TN groups in the selected variables. The null hypothesis was rejected for all the prognostic variables. Therefore their statistical distributions should be considered to be different and the differences to be significant (see Appendix).

Additionally the Maximum Percentile Distance method also found a subset of variables with lower accuracy (78%): MBC, ALC, ZAP70, WBC and B2M. Moreover, the results using the minimum class Entropy were quite similar (76.8%): ZAP70, BU, WBC, ALC and MBC.

CT is recommended in patients with advanced and progressive disease. Thus, the amount of malignant leukemia cells that it is measured by the different counts of leucocytes; particularly WBC (White Blood Cells count), ALC (Absolute Lymphocyte Count) and MBC (Monoclonal B Cell Count) are key clinical parameters. Nevertheless, these variables are not currently used to select patients who may benefit from CT. On the other hand, AGE, B2M and ZAP70 are traditional clinical parameters that have demonstrated their prognostic importance independently of the clinical stage. Our results also indicated the great prognostic significance of other variables that are mainly related with the characteristics of the immune system and are not currently used as prognostic markers in this disease. The fact that the prediction accuracy is barely above 80% means that these variables only contain partial information to establish the need of CT and/or to incorrect medical decisions that might input noise in the class assignment.

Table 2 shows the median/mean signatures for the 4 groups of the confusion matrix for the main decision variables found by this methodology. We can observe that there exists a significant distance between the mean signatures of the TP and TN

groups, being the median/mean signatures in all the decision variables much higher in the TP group. Moreover, the distance between the median and the mean values of the decision variable distributions is much higher in the TP and in the FP groups, meaning a higher variability in these groups:

- The normal value of B2M is less than 2 mg/L [26]. Levels of B2M can be elevated in multiple myeloma and lymphoma. Besides, elevated values (>4 mg/L) are known to be an indicator of poor prognosis and survival [27]. In our case B2M is higher than this cut-off value (4.24) for the patients in the TP group.
- For the second decision variable, the normal value of WBC in the blood is 4.5-10.0 Kcells/microL. In our case the patients of the TN group have a mean WBC value (16.8 Kcells/microL) that exceeds four times the minimum normal value. Also the patients in the TP group show even higher mean WBC values (61.8 Kcells/microL).
- The reference range for the ALC is 4.5-11.0 Kcells/microL. It can be also observed that the ALC mean value in the TP group (47.6 Kcells/microL) exceeds 4 times the maximum normal value.
- Finally, the MBC is also very high (40.3 Kcells/microL) in the TP group compared to the TN group (8.4 Kcells/microL). The definition of CLL implies having a rate of CLL-phenotype B-cell lymphocytes higher than 5 Kcells/microL.

This analysis shows the typical profile of CLL patients with need of CT. The same tendencies are observed for the corresponding median values.

With respect to the analysis of the classification errors, the mean signatures of the FN group (patients that need CT and are incorrectly predicted) are very close to the mean signatures of the TN group. These patients will never be correctly predicted according to this classifier. The mean and median signatures of the FP group have the following singularities:

1. The mean B2M value (4.58 mg/L) is even higher than the corresponding B2M mean value in the TP group (4.24 mg/L). The same is observed for the median values.
2. Their mean WBC, ALC and MBC values are closer to the corresponding

mean values of the TN group, exceeding in all the cases the mean values of the TN group. These differences are smaller in the case of the median values. These patients could be detected using only these three variables, not considering the value of B2M in these patients that is distorting the prediction.

Furthermore, to understand the ambiguity in the CT prediction, it should be taken into account that the criteria used to establish the need of CT [25] sometimes have not correlation with the biological data. The reason is that some patients are diagnosed in early stages of the disease when a low burden tumor mass has been detected but they have a very fast progression which implies the need of CT.

3.2. Autoimmune Disease development

An autoimmune disease (AD) occurs when an adaptive immune response is mounted against self-antigen. In CLL, an autoimmune response against red blood cells (known as autoimmune haemolytic anemia), and an autoimmune response against platelets (known as immune thrombocytopenia) are severe complication of this disease. To the best of our knowledge no prognostic factors capable to predict the presence or development of an autoimmune disease in CLL patients have been currently disclosed. In our cohort only 16 patients (out of 263, therefore there are 2 missing values since the total cohort is 265) have shown autoimmune disorders. Therefore this classification problem, independently of the data sampling, is intrinsically highly unbalanced.

The shortest list of prognostic variables with the highest accuracy (97.3%) was found by the Fisher's ratio method and includes 13 clinical variables: PLT, RET, ALB, HGB, BU, UR, MCV, NCC, K, WBC, LDH, ALC and MBC. Furthermore, considering only the first nine attributes the predictive accuracy was 95.4%. Besides, only the two first attributes provided a predictive accuracy of 91%. Figure 3 shows the ROC and the Recall (or True Positive Rate -TPR) against Precision (or Positive Predicted Value - PPV) curves throughout all possible probability thresholds for the AD classification problem. The optimum result ($p_{th} = 0.5$) shows that 62.5% (TPR) of the patients that have AD and 99.6% (True Negative Rate or Specificity -SPC) of the patients that do not have AD are correctly predicted. Moreover, over that probability threshold we get a Precision (or Positive Predicted Value – PPV) of 90.1%. However,

other probability thresholds could be adopted depending on the Recall/Specificity balance, and therefore on the PPV as well. The False Discovery Rate (FDR) in this case is 9.1%. The confusion matrix is the following one:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 10 & 1 \\ 6 & 246 \end{pmatrix}.$$

The True Positives (TP) group is formed in this case by the patients that present AD (+) and are correctly predicted and True Negatives (TN) correspond to the patients that do not have AD (-) and are correctly predicted. Similarly, the False Positives (FP) are the patients that do not have AD (-) and are not correctly predicted and the False Negatives (FN) correspond to the patients that present AD (+) and are not correctly predicted. As in the previous section, we have performed the T-test and the Mann-Whitney U-test to analyze the differences between the TP and TN groups. The null hypothesis was rejected for all selected variables, except for K and LDH in the T-test (see Appendix).

Additionally the percentile distance method also found a subset of variables with 95.1% accuracy composed only by one prognostic variable: NCC. Entropy method also found a subset of 4 prognostic variables with 94.3% accuracy: TLC, T8C, NCC and MBC. PLT and RET, that were ranked in the first positions by the Fisher's Ratio, were found by the Entropy method in the fifth and sixth positions (TLC, T8C, NCC, MBC, RET and PLT), but the accuracy of this final subset was 93.2%.

PLT and RET appears in the first two positions of the FR list. They are responsible for most of the discriminatory power of the reduced base of features and the rest of variables span high frequency details in the classification. They also appear in the first positions of the list using Entropy method. It seems they could have an important role in the development of an autoimmune disease. Table 3 shows the medians and means for the 13 prognostic variables for the 4 groups of the confusion matrix. The differences between the means in TP and TN groups decrease with the Fisher's ratio. Prognostic variables with lower Fisher's ratios (secondary variables) also contribute to improve the discrimination. Except for the main variable, PLT, and the secondary variables HGB and K, the mean and median values are higher in the group with autoimmune disease (TP). The analysis of the two main prognostic variables shows that patients that develop AD and are correctly predicted (TP) have much lower medians and means PLT values (97.7/95.0 Kcells/microL). The normal

platelet count lays in the range 150-450 Kcells/microL, being the average 237 Kcells/microL in men, and 266 in women. On the other hand, the reticulocyte count (RET) in the TP group almost doubles (136 Kcells/microL) the average RET count in patients with no AD (70 Kcells/microL). Median values also show similar tendencies.

The False Positives (FP group) is composed in this case only by 1 sample, whose signature is closer for all the 13 variables to the TP group, except for PLT, RET that are somewhere in between the median/mean values for TP and TN. This fact points out the difficulty of classifying this sample, and it can be concluded that it could be a 'biological' outlier. On the other hand, the FN group is composed by 6 samples. The mean PLT count (147 Kcells/microL) of the FN group lies between the mean value for the TP (95 Kcells/microL) and TN (202.2 Kcells/microL) groups. The RET count is however closer to the TN group showing a tendency to very low median values (54.4 Kcells/microL).

The percentile distance method found a subset of variables with 95.1% accuracy composed only by the Natural killer Cell Count (NCC). The mean NCC value in the TP group (2251 cells/microL) is higher than in the TN (741 cells/microL) and FN (393 cells/microL) groups. Natural killer cells provide rapid responses to virally infected cells and respond to tumor formation. Therefore, this result suggests a possible link between AD development, viral infection and tumor progression. The percentile method also gives a great importance to IgM due to the higher values in the group of patients without AD (TN group with a mean of 1.12 g/L) with respect to the TP group (mean value of 0.36 g/L). This result is important since IgM is the first antibody to appear in response to initial exposure to antigens [28] and lower levels of this inmunglobulin is related to selective immunoglobulin M deficiency, which in turn is also related with autoimmune disorders like celiac disease or systemic lupus erythematosus [29].

Overall, these results show the importance of variables associated with the characteristics of platelets and red cells, which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia, such as PLT, HGB, MCV and RET. Other variables depend on the presence of autoantibodies (COOMBS) or products or symptoms derived from the lysis of blood cells (BU, LDH and SMG). Moreover, some variables associated with the immunological characteristics of patients, such as IgM, IgG, IgA, TLC, NCC and T8C, constitute a relevant subset of variables that may predict an autoimmune disease occurrence. The association of

these variables with an autoimmune disease is not unexpected based on the biology of CLL, but we would like to highlight that no prognostic factors or system may currently predict the development of an autoimmune disease in the clinical practice. To the best of our knowledge this is the first description so far that a group of clinical variables obtained at diagnosis of CLL patients may predict an occurrence of an autoimmune disease.

3.3. Summary of the results

Finally, Table 4 summarizes the main results found for both classification problems (CT and AD): the optimum reduced set of features, the LOOCV accuracy, the hold out (HO) mean accuracy over 100 different random simulations using 75% and 25% of samples for training and testing, the Sensitivity or True Positive Rate (TPR), and the Specificity or True Negative Rate (SPC) statistics. TPR and SPC values are important due to the impact on the patients of the decision taken by physicians.

It is possible to observe that:

1. The median accuracy of the predictions is quite stable with respect to the LOOCV accuracy.
2. The TPR/SPC statistics are optimally balanced in all the problems. The TPR/SPC statistics might be the target of a different optimization for the weights of the linear classifier depending on the risk that is given by the medical doctors to the False Positives (FP) and False Negatives (FN) diagnostic in each classification problem. This approach has been adopted to predict response to treatment in Hodgkin Lymphoma [17].

4. CONCLUSIONS

Different prognostic factors are presented in this paper to predict two clinically important classification problems for CLL patients: Chemotherapy Treatment assessment and Autoimmune Disease development.

From the machine learning point of view, working imputed data produced better results in reliability (accuracy) than working with raw data. Fisher's ratio and percentile distance are the feature selection methods that produced the best biomarkers in terms of medical interpretability. The minimum-size of variables is

established using BFE. The class prediction is based on a simple classifier, and its accuracy is determined by LOOCV experiment. The results show that the accuracies are rather high and the difference between both experiments LOOCV and 100 repetitions of a Hold Out (75/25) is quite low, which highlights the robustness of the methodology. In addition, risk assessment ROC curves are provided for each problem and show a good balance between False Positives and False Negatives.

From a medical point of view, machine learning methods allow the identification of clinical variables obtained at diagnosis of CLL patients, which may predict the development of AD and the need of CT. These variables are obtained at diagnosis of CLL patients on a regular basis, and consequently, their use does not increase the cost or complexity of the diagnosis in CLL patients. The need of CT seems to be related to the amount of malignant leukemia cells that are measured by the different leucocytes counts.

The best prognostic variables to predict the need of CT were B2M, WBC, ALC and MBC. Although the results concerning these prognostic variables are well known in other plasma disorders, this analysis served to conclude that these variables only carry partial information to adopt this important decision, that most of the times, is taken based on criteria that have not correlation with the biological data. To the best of our knowledge this is the first description so far that a group of clinical variables obtained at diagnosis of CLL patients may predict an occurrence of an AD, which is a severe and currently unpredictable complication of this disease. These results show the importance of variables associated with the characteristics of platelets, reticulocytes and natural killers (PLT, RET and NCC), which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia. Additionally, machine learning methods focus on the relevance of some variables, such as the immunological ones, which may have an important impact on the prognosis of CLL patients, but they are not currently used by hematologists. Particularly, this analysis has shown that the low sampling frequency of RET and ZAP-70 could be troubling given their predictive significance in all the problems that have been treated: RET is a key factor for predicting AD, whilst ZAP-70 seems to be important for predicting the need of CT.

In conclusion, machine learning methods allow an accurate prediction of risk in CLL related problems. Additionally, they may establish the relevance of clinical variables that are not widely used as prognostic factor in this disease. The prognostic

significance of these variables may probably reflect the relevance of some clinical aspects of this disease that are more important for prognosis than it is currently thought. This bioinformatics system can be easily applied in medical practice and updated along time through a simple computer program or excel spreadsheet (see supplementary material file **"**CLL_predictor.xls").

Acknowledgments

Enrique J. de Andrés was supported by the Spanish Ministerio de Economía y Competitividad (grant TIN2011-23558). The medical analysis was supported by the Fondo de Investigaciones Sanitarias (Instituto Carlos III-grant PI12/01280). No other financial support has been received to perform this retrospective analysis. We would like to acknowledge Dr. Stephen T. Sonis for his constructive review and suggestions that served to improve the translational approach shown in this manuscript.

Ethics statement

This study was approved by the Ethics Committee of Clinical Investigation of Principado de Asturias (date: 21th of January of 2009; nº1/2009). All the patients signed an informed consent to participate in this study with the approval of the Ethics Committee. All studies were performed in accordance with the ethical standards of the Declaration of Helsinki and informed consent was obtained from all patients and controls.

Conflict of interest

No conflict of interest exists for the authors of this paper.

Authors' contributions

JLFM and EJAG prepared the data, designed the machine learning methodology, carried out the experiment, analyzed and interpreted the results and drafted the manuscript. OLR and JJC revised the design of the methodology critically, analyzed the results and drafted the manuscript. LHZ and AAH participated in the acquisition of the data, analyzed and interpreted the results and drafted the manuscript. SG and APG participated in the acquisition of the data, analyzed and interpreted the results, established main clinical conclusions and drafted the manuscript. All authors read and approved the final manuscript.

References

[1]. Hallek M, Wanders L, Ostwald M, et al. Serum beta(2)-microglobulin and serum thymidine kinase are independent predictors of progression-free survival in chronic lymphocytic leukemia and immunocytoma. *Leuk Lymphoma* 1996; 22(5-6): 439–47.

[2]. Zenz T, Mertens D, Dohner H, Stilgenbauer S. Molecular diagnostics in chronic lymphocytic leukemia -pathogenetic and clinical implications. *Leuk Lymphoma* 2008; 49(5):864–73.

[3]. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated ig v(h) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* 1999; 94(6): 1848–54.

[4]. Crespo M, Bosch F, Villamor N, et al. Zap-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. *N Engl J Med* 2003; 348(18): 1764–75.

[5]. Damle RN, Wasil T, Fais F, et al. IgV gene mutation status and cd38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 1999; 94(6): 1840–7.

[6]. Gonzalez-Rodriguez AP, Contesti J, Huergo-Zapico L, et al. Prognostic significance of cd8 and cd4 t cells in chronic lymphocytic leukemia. *Leuk Lymphoma* 2010; 51(10): 1829–36.

[7]. Rai KR, Sawitsky A, Cronkite EP, Chanana AD, Levy RN, Pasternack BS. Clinical staging of chronic lymphocytic leukemia. *Blood* 1975; 46(2): 219–34.

[8]. Binet JL, Auquier A, Dighiero G, et al. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer* 1981; 48(1): 198–206.

[9]. Olden JD, Lawler JJ, Poff NL. Machine learning methods without tears: a primer for ecologists. *Q Rev Biol* 2008; 83(2): 171–93.

[10]. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform* 2008; 77(2): 81–97.

[11]. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* 2006; 2:59–77.

[12]. Troyanskaya OG, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; 17(6): 520–5.

[13]. Fisher RA. The use of multiple measurements in taxonomic problems. *Annals*

of Eugenics 1936; 7(7): 179–88.

[14]. Yang F, Mao K. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2011; 8(4): 1080–92.

[15]. Shannon C. A mathematical theory of communication. *Bell System Technical Journal* 1948; 27: 379–423, 623.

[16]. Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1993.

[17]. deAndrés-Galiana EJ, Fernández-Martínez JL, Luaces O, et al. On the prediction of Hodgkin Lymphoma treatment response. *Clin Transl Oncol* 2015; 17(8): 612-9.

[18]. Saeys Y, Inza In, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; 23(19): 2507–17.

[19]. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002; 46(1-3): 389–422.

[20]. Swets JA. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers. By JOHN A. SWETS. Mahwah, NJ: Lawrence Erlbaum Associates; 1996.

[21]. Fernández-Martínez JL, Fernández-Muñiz Z, Tompkins MJ. On the topography of the cost functional in linear and nonlinear inverse problems. *Geophysics* 2012; 77(1):1-15.

[22]. Fernández-Martínez JL, Fernández Muñiz Z, Pallero G, Pedruelo González LM. From Thomas Bayes to Albert Tarantola. New insights to understand uncertainty in inverse problems from a deterministic point of view. *J App Geophys* 2013; 98: 62-72.

[23]. Fernández-Martínez JL, deAndrés-Galiana EJ, Sonis S. Design of Biomedical Robots for the Analysis of Cancer, Neurodegenerative and Rare Diseases. *Proceedings of the International Conference on Man-Machine Interactions* 2015; 391(4): 29-44.

[24]. Saligan L, Fernández-Martínez JL, deAndrés-Galiana EJ, Sonis S. Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer. *Cancer Informatics*, 2014; 13:141–152, 12.

[25]. Hallek M, Cheson BD, Catovsky D, et al. Guidelines for the diagnosis and

treatment of chronic lymphocytic leukemia: a report from the international workshop on chronic lymphocytic leukemia updating the national cancer institute-working group 1996 guidelines. *Blood* 2008; 111(12): 5446–56.

[26]. Pignone M, Nicoll D, McPhee SJ. Pocket guide to diagnostic tests (4th ed.). New York: McGraw-Hill; 2004.

[27]. Munshi NC, Longo DL, Anderson KC. Plasma Cell Disorders. In Loscalzo J, Longo DL, Fauci AS, Dennis LK, Hauser SL. Harrison's Principles of Internal Medicine. McGraw-Hill Professional; 2011.

[28]. Houghton Mifflin Company. Immunoglobulin M. The American Heritage Dictionary of the English Language; Fourth Edition 2004.

[29]. Yel L, Ramanuja S, Gupta S. Clinical and Immunological Features in IgM Deficiency. *Int Arch Allergy Immunol* 2009; 150 (3): 291–8.

LIST OF CAPTIONS

Figure 1: Methodology flowchart.

Figure 2: A) ROC curve. B) Sensitivity (or True Positive Rate -TPR) and Precision (or Positive Predicted Value - PPV) for Chemotherapy Treatment. The optimum result (TPR = 63.4 and PPV = 64.3) is obtained for $p_{th} = 0.47$.

Figure 3: A) ROC curve. B) Sensitivity (or True Positive Rate -TPR) and Precision (or Positive Predicted Value - PPV) for Autoimmune Disease occurrence. The optimum result (TPR = 62.5 and PPV = 90.1) is obtained for $p_{th} = 0.5$.