Brief Report

Machine Learning Provides an Accurate Classification of Diffuse Large B-Cell Lymphoma from Immunohistochemical Data

Carlos Bruno Tavares Da Costa[1]

[1]Hematology Unit, Department of Medicine, Hospital das Forças Armadas, Lisbon, Portugal

Received: 8 March 2018 Accepted: 8 May 2018 Published: 13 June 2018

Abstract

Background: The classification of diffuse large B-cell lymphomas into Germinal Center (GCB) and non-GC subtypes defines disease subgroups which are different both in terms of gene expression and prognosis. Given their clinical significance, several classification algorithms have been designed, some by making use of widely availability immunohistochemical techniques. Despite their high concordance with gene expression profiles (GEP) and prognostic value, these algorithms were based on technical and biological assumptions that could be improved in terms of performance for classification. **Methods:** In order to overcome this limitation, a new algorithm was obtained by analyzing a previously published dataset of 475 patients by using an automatic classification tree method. **Results:** The resulting algorithm classifies correctly 91.6% of the cases when compared to GEP, displaying a Receiver-Operator Characteristic (ROC) area under the curve of 0.934. Noteworthy features of this algorithm include the capability to classify GEP-unclassifiable cases and a significant prognostic value, both in terms of overall survival (60 months for non-GC vs not reached for GCB, $P = 0.007$) and progression-free survival (61.9 months vs not reached, $P = 0.017$). **Conclusion:** By using a machine learning classification method that avoids most pre-assumptions, the novel algorithm obtained is accurate and maintains relevant features for clinical implementation.

Keywords: Cell of origin, immunohistochemistry, lymphoma, machine learning, prognostic

Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin's lymphoma, accounting for approximately 30% of all cases.[1] The improvement in the prognosis of patients observed over the past two decades was due not only to increasingly effective therapies but also to clearer definitions of disease and prognostic factors, stemming from more robust diagnostic and staging techniques. One of the most significant discoveries was the definition of distinct disease entities based on gene expression patterns. As originally described by Alizadeh, DLBCL can be divided into subgroups with germinal center B-cell (GCB)-like, activated B-cell (ABC)-like, and type 3 gene expression profiles (GEPs), adding fundamental information to the previously suspected biological diversity of this disease.[2] In fact, these subgroups differ by the expression of more than 1000 genes, which is comparable to the difference between acute lymphoid and myeloid leukemias.[3] More importantly, this model defines subgroups with different prognoses, where the ABC gene signature is an independent adverse prognostic factor, even in the era of combination therapy with Rituximab, Cyclophosphamide, Doxorubicin, Vincristine and Prednisolone (R-CHOP).[2,4,5]

Despite the implementation of GEP into clinical practice in recent years, DNA- and RNA-based prognostic methods remain expensive and technically challenging. Thus, alternative and simpler IHC techniques have been explored. Among the earliest described, the Hans algorithm allowed

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_14_18

Address for correspondence: Dr. Carlos Bruno Tavares Da Costa, Hematology Unit, Department of Medicine, Hospital das Forças Armadas, Lisbon, Portugal.
E-mail: carlosbrunotcosta@gmail.com

How to cite this article: Costa CB. Machine learning provides an accurate classification of diffuse large b-cell lymphoma from immunohistochemical Data. J Pathol Inform 2018;9:21.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2018/9/1/21/234448

the distinction between GCB and non-GC DLBCL subtypes using a set of measurable proteins that included CD10, BCL-6, and MUM1/IRF4.[6] It retained prognostic significance for patients treated with CHOP, but the concordance with GEP was only 71% for GCB and 88% for non-GC lymphomas. More importantly, it was developed in the prerituximab era and its application to patients treated with R-CHOP led to variable results and may have lost its prognostic value in this setting.[7,8] Subsequently, several other algorithms based on immunohistochemical (IHC) stains and tissue microarray techniques were developed to overcome these limitations.[4,5,9,10] These new, more accurate, algorithms have introduced new proteins into the set of relevant attributes, with significant discriminant and prognostic powers (FOXP1, GCET1, and LMO2). The rationale used to design these algorithms was mostly based on trial and error approaches, technical considerations related to tissue staining, and a certain number of biological assumptions that, despite their relevance, lack mathematical validity. One could take as an example the Visco-Young algorithm, a 3-marker signature that uses CD10, FOXP1, and BCL-6. It has a 92.6% concordance with GEP and retains a strong independent prognostic value in patients treated with R-CHOP.[4] These authors argue that CD10 should have a prominent role because it is part of the initial diagnostic staining panel in most practices and shows the best concordance in different studies performed in different laboratories. Oppositely, a minor role is given to BCL-6 due to the high variability obtained in staining, which might be related both to the avidity of the antibodies used in this IHC analysis and the natural variability of this epitope.[11] The remaining algorithm transposes the known B-cell maturation steps at the germinal center to the classification flow. Despite their clinical validity, these technical and biological assumptions may introduce a significant bias into the rationale of the classifier and lead to an under- or overestimation of the role of each of these markers. Moreover, these methods do not account for aberrant differentiation pathways which one could expect to encounter in such diseases. Furthermore, in the Visco-Young algorithm, the cutoff values for each marker were determined using the Youden index from receiver-operator characteristic (ROC) curves determined for each marker individually. Although commonly used, this method might fail to provide a robust basis for the multivariate nature of the proposed algorithm. Moreover, as others before them and for practical reasons, the authors deliberately used cutoffs for GCET1 and FOXP1 that were different from those given by the Youden index. Still, they found out that this did not change the sensitivity and specificity of these markers. While this may be valid for each marker individually, it may introduce a significant bias into a multifactor analysis, which highlights the relevance of an independent validation.

Machine learning comprises a group of techniques that allow the use of large, complex datasets to build classification models and can provide the basis to address the issues identified above. Several data mining and analysis software packages are available for academic use, each comprising several algorithms for data analysis. For this classification problem, the WEKA (University of Waikato, New Zealand) software and the C4.5 statistical classifier were used. This algorithm is based on the concept of information entropy and generates decision trees in a recursive way, where each node splits data as effectively as possible in terms of enrichment of the resulting branches in any one of the categories being studied.

In this article, I present a DLBCL IHC classification algorithm obtained through machine learning classification methods whereby ignoring any pre-assumptions (beyond the limited set of markers available), I expect to provide additional validity to the results that emerge and most importantly, raise new hypotheses.

Methods

Patient data were obtained from the Visco and Young dataset, which is kindly available as supplementary materials to their original article.[4] The data of all 475 patients used in the design of their algorithm were also used in this work. These data were processed and analyzed using the machine learning WEKA package, v. 3.6.11. To design the new algorithm, the GEP-unclassifiable (UC) cases were removed from the dataset. For the remaining 431 cases, a new algorithm was obtained using the J48 classification method, a derivation of the C4.5 method implemented in WEKA.[12] To obtain a simplified classification tree with significant groups, a minimum of 20 cases were imposed into each class. A ten-fold cross-validation method was used. The resulting classification tree was then applied to the entire original dataset including GEP-UC cases. To evaluate the performance of the classifier, ROC curves were obtained and the area under the curve was used as an overall measure of sensitivity and specificity. Survival curves for GCB and non-GC groups were obtained using the Kaplan–Meier method and compared by the log-rank test. The classification results were included in a Cox proportional hazards model for multivariate analysis for prognosis. The level of significance used to justify a statistically significant effect was 0.05.

Results

The resulting algorithm is shown in Figure 1. Similarly, to the Visco-Young algorithm, it includes CD10 as stem marker, and subsequently, the classification tree branches out to include MUM1, FOXP1, and BCL-6 markers. The performance of this algorithm in terms of efficacy of classification is similar to that of the Visco-Young and Choi methods and superior to the Hans method: 395 cases were correctly classified, achieving a 95.7% true positive rate for GCB and 87% true positive rate for ABC, with overall 91.6% correctly classified cases. This corresponds to a kappa statistic of 0.83 and a ROC area under the curve of 0.934. A total of 231 cases were classified as GCB and 200 cases were classified as non-GC. The vast majority of misclassified cases have a ABC GEP (21/27 cases) but entered the IHC category for GCB.

By applying the new classification to the original clinical data, it was possible to compare progression-free survival (PFS) and overall survival (OS) among GCB and non-GC patients as classified by the new algorithm. The Kaplan–Meier plots shown in Figure 2 exclude GEP-UC patients and demonstrate significant statistical differences among subgroups ($P = 0.024$ and $P = 0.017$ for PFS and OS, respectively).

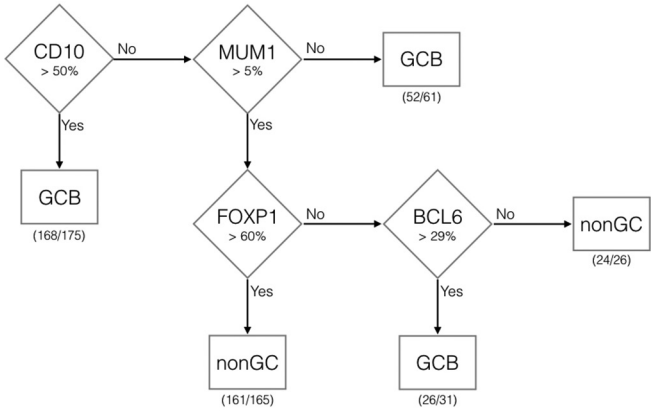

Figure 1: Algorithm for immunohistochemical classification obtained by applying a classification tree method to the Visco-Young dataset after removing the unclassifiable cases. The numbers below the boxes indicate the number of cases correctly classified and the total number of cases classified in the category identified in the corresponding box. GCB: germinal center B-cell, non-GC: non-germinal center

One of the potential advantages of IHC methods is the ability to classify the previously GEP-UC cases and derive prognostic information from them. Figure 3 depicts the Kaplan–Meier plots obtained after applying the new classification algorithm to all the cases, including GEP-UC. Median OS and PFS are, respectively, 60 and 61.2 months for non-GC and not reached for GCB subgroups. Differences between subgroups are statistically significant and because these results were maintained after inclusion of the GEP-UC patients, suggest that this algorithm might have intrinsic prognostic properties, independently of the correlation with GEP.

When analyzing the correspondence between the new algorithm and the Visco-Young in terms of classification of GEP-UC cases (n = 44), only one out of 21 GCB cases are not a match, but among non-GC cases, 6 out of 23 (26%) cases disagree in terms of IHC classification [Figure 4]. This relates to the fact that a higher proportion of GEP-UC cases are classified as GCB by the new algorithm (26/44) compared to the Visco-Young algorithm (21/44). The reclassification did not have an apparent impact on survival.

On a multivariate analysis, an IPI score of 3 or more, lack of complete response, non-GC class as predicted class by the new classifier, and LDH above normal were significantly associated with worse survival. Both gender and a poor performance status did not reach statistical significance. These results are summarized in Table 1.

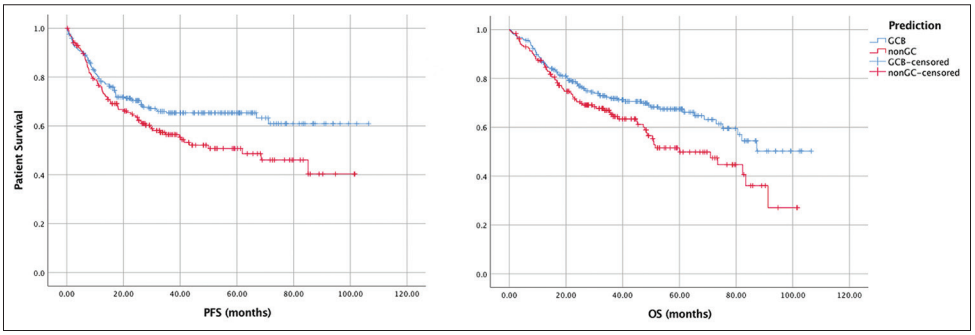

Figure 2: PFS and OS among GCB and non-GC patients, as classified by the new algorithm, excluding GEP-unclassifiable cases. PFS: $P = 0.024$ (median 61.9 months for non-GC vs. not reached). OS: $P = 0.017$ (median 60 months for non-GC vs. not reached). GCB: germinal center B-cell, non-GC: non-germinal center, OS: Overall survival, PFS: progression-free survival

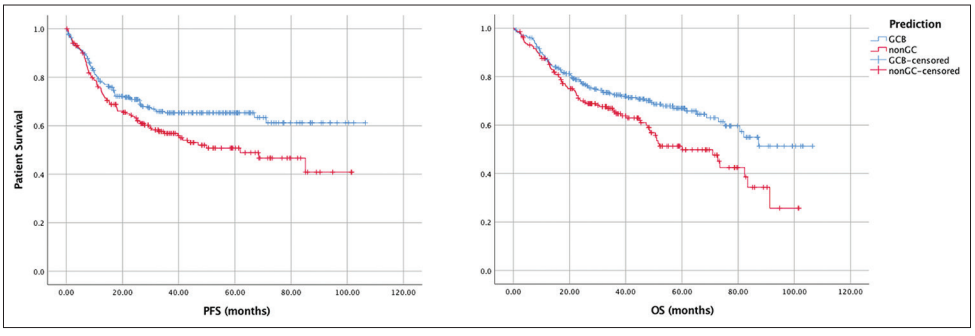

Figure 3: PFS and OS among GCB and non-GC patients, as classified by the new algorithm, including GEP-unclassifiable cases. PFS: $P = 0.017$ (median 61.9 months for non-GC vs. not reached). OS: $P = 0.007$ (median 60 months for non-GC vs. not reached). PFS: progression-free survival, OS: Overall survival

Discussion

The new algorithm described was obtained by applying the machine learning method J48, a derivation of C4.5, to the Visco-Young dataset and includes CD10, MUM1, FOXP1, and BCL-6 into an IHC classification tree that can be applied to DLBC lymphoma. Reference algorithms and their respective concordance with GEP are summarized in Table 2. No other previously described algorithm uses such a combination of markers and cutoffs. Another relevant aspect of this result, unlike previous algorithms, is that it was obtained without any *a priori* assumptions. This is most relevant because it avoids technical preconceptions (such as in the case of BCL-6, as discussed in the introduction, but also CD10) and methodological shortcuts (as in the case of the simplified cutoff values, also discussed in the introduction), both of which may introduce bias into a classifier. Moreover, this method avoids any kind of biological presumptions which, from a more conceptual perspective is very relevant, as recapitulating the physiological differentiation pathways to explain malignant phenotypes may fail to take into account aberrant pathways involved in pathological states. Interestingly, both the stem and the terminal markers of this new algorithm coincide with the Visco-Young algorithm, aiding to the validation of some of the original biological assumptions. Furthermore, FOXP1 is placed immediately before BCL-6 in the classification flow, again adding to the validity of the original biological rationale used by this and other previous algorithms where the steps of B-cell maturation are recalled. However, this algorithm discards GCET1 (used in both the 4-marker Visco-Young, Tally and Choi algorithms) and uses MUM1 at a 5% cutoff point. Below this value, most cases are classified as GCB (43/52, 82.6%), which implies that even small expression levels of MUM1 are significantly associated with a non-GC-like phenotype. Furthermore, the proposed algorithm indicates that most of MUM1 positive cases have high levels of FOXP1, and both are associated with the non-GC phenotype. This may indicate a biological rationale for an interplay between the pathways involved, but such inference is out of the scope of this work, despite the known role of PPAR-alpha in the expression of both FOXP1 and MUM1.[13]

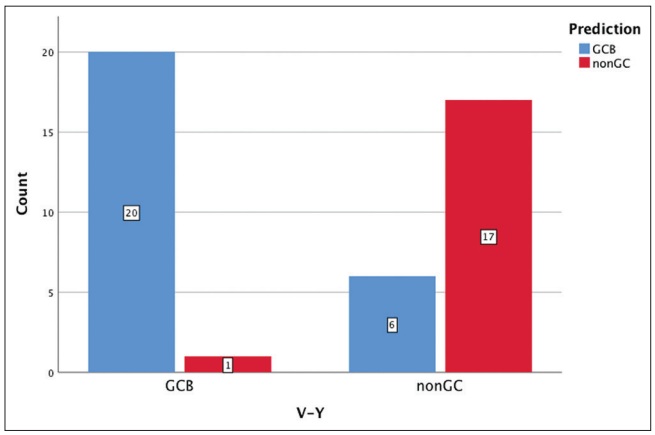

Figure 4: Correspondence between the Visco-Young and the new algorithm described ("prediction") when applied exclusively to GEP-unclassifiable cases. GEP: Gene expression profiles

Compared to GEP, this new classification algorithm correctly classified 91.6% cases. Most of the misclassified cases have an ABC GEP. This is contrary to other previous analyses, where the proportion of misclassified cases by IHC compared with GEP was higher when defining the GCB subtype.[5,14] Although the reason for these differences remains elusive, one possible explanation could be the generally poor performance of BCL-6 as a GCB marker.[11] High levels of BCL-6 mRNA are associated with a good prognosis, but BCL-6 staining and analysis is known to be highly variable and has a poor correlation with mRNA levels.[15] In this work, BCL-6 ≥30% is associated with a GCB-like phenotype in a minority of the MUM1+ cases with lower levels of FOXP1+. This suggests that BCL-6 could be a weaker marker of phenotype compared to FOXP1 and MUM1, which might also help explain the variable results obtained earlier in terms of the prognostic significance of the BCL-6 immunophenotype.[16,17] Previously, different algorithms suggest diverse weights for BCL-6 in terms of the number of patients discriminated according to this marker, and it has been shown to be associated with both GCB and non-GC markers.[18] If BCL-6 is removed from the dataset, a new algorithm can be derived, which is similar to the one presented above but incorporates GCET1 in the

Table 1: Multivariate analysis of risk factors for progression-free survival and overall survival

	PFS				OS			
	Significant	HR	95.0% CI for HR		Significant	HR	95.0% CI for HR	
			Lower	Upper			Lower	Upper
Gender	0.067	1.337	0.98	1.824	0.53	1.109	0.803	1.531
ECOG ≥3	0.868	0.958	0.579	1.586	0.268	0.758	0.463	1.238
IPI high	0.002	0.568	0.399	0.809	0.003	0.575	0.4	0.827
CR	0	5.853	4.201	8.157	0	6.596	4.667	9.324
Prediction	0.024	0.691	0.501	0.952	0.011	0.653	0.469	0.908
LDH	0.078	0.839	0.69	1.02	0.047	0.813	0.662	0.997

ECOG≥3: ECOG performance status grades 3 or 4, PFS: Progression-free survival, OS: Overall survival, HR: Hazard ratio, CI: Confidence interval, IPI high: International Prognostic Index of 4 or 5, CR: Complete response, Prediction: Predicted class by the proposed model, LDH: Lactate dehydrogenase, ECOG: Eastern Cooperative Oncology Group

Table 2: Reference algorithms and their respective concordance with gene expression profiling

Algorithm	Components (and cut-offs)	Concordance with GEP (%)
Hans (7)	CD10 (30%), BCL-6 (30%), MUM-1 (30%)	80.8
Choi (9)	GCET1 (80%), MUM1 (80%), CD10 (30%), BCL-6 (30%), FOXP1 (80%)	93
Visco-Young (4)	CD10 (30%), FOXP1 (60%), BCL-6 (30%)	92.6
Costa	CD10 (50%), MUM1 (5%), FOXP1 (60%), BCL-6 (29%)	91.6

GEP: Gene expression profiles

same terminal position, at a 5% cutoff point. This model has a worse discriminant capacity, with only 378 (87%) correctly classified cases. The reason why GCET1 was not included in the algorithm now described is that its role as a classifier superimposes that of BCL-6 but with an inferior accuracy, making the algorithm less robust.

If the algorithm presented were to be applied using the 30% cutoff for every input variable (as done in the Hans model), the accuracy would fall to 86.5%. This illustrates that input order and weighting are important aspects of this and every other algorithm. However, the most relevant feature of an algorithm such as this is expressed by its prognostic significance in terms of OS and PFS, even for the GEP-UC cases. This adds to the validity of this algorithm, which seems to have prognostic capabilities beyond its correspondence to GEP. This is an important point in favor of IHC methods, which, despite their variable results, do not leave any cases unclassified while retaining independent significance in multivariate analysis for both OS and PFS.

Comparing with the Visco-Young algorithm, the cutoff values obtained with the new algorithm are similar for BCL-6 (30%) and FOXP1 (60%) but different for CD10 (50% vs. 30%) and for MUM1 (5% vs. 30%). Both FOXP1 and MUM1 were used also in the Choi algorithm (both at 80% cut-off point), and Hans described the use of MUM1, CD10, and BCL-6 at the 30% cutoff value. The differences among these and other previously published algorithms stems from variable staining performances, different visual analyses and demonstrates the disparate interpretations of apparently similar data. This calls for the development of robust staining methods as well as mathematical methods that can define an accurate, valid algorithm. I believe that machine learning methods fulfill part of this task and could provide a generalizable method to approach new classification efforts. The relatively superior performance of this new algorithm in terms of classification of GCB (vs. non-GC) lymphomas is an interesting feature with potential utility in trials dedicated to this subgroup of DLBCL. The similarities between the results described here and the Visco-Young algorithm validates it in terms of clinical utility and biological logic. Future analysis could make use other classification methods. Beyond J48, Bayesian methods were also experimented, but the results were not superior to J48 in terms of accuracy (data not shown).

Conclusion

In this article, a new DLBCL classification model based on IHC stains is described, obtained using the machine learning algorithm J48 that has both high correspondence to GEP and prognostic significance. This is a novel approach to the IHC classification of lymphomas. Unlike prior models, the new algorithm lacks any *a priori* biological assumptions (beyond the limited set of markers), but its similarities to some of the prior models seem to support those. This includes the potential association between MUM1 and FOXP1 in non-GC cells and the minor role of BCL-6 in GCB cases. In a multivariate analysis, the new model is an independent predictor of survival. Compared to gene expression profiling, IHC algorithms such as the one described here have an easier, cheaper implementation, and allow clinicians to classify GEP-UC cases and derive prognostic information from them. In our case, the prognostic significance of the algorithm was shown to be preserved after the inclusion of the UC cases, which means that the model has prognostic power beyond the correspondence with GEP. Future work might aim at developing an IHC model that maximizes prognostic performance. The use of machine learning algorithms provides robust tools to process large amounts of data and define new classification models. The consequent clinical and biological insights obtained should be further explored and validated. Using these methods in other contexts may offer novel insights into the biological foundations of disease and drive future research.

Acknowledgments

I would like to publicly acknowledge the University of Waikato, namely Professor Ian H. Witten for developing WEKA and providing it as a valuable, robust, open-source tool.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Martelli M, Ferreri AJ, Agostinelli C, Di Rocco A, Pfreundschuh M, Pileri SA, *et al.* Diffuse large B-cell lymphoma. Crit Rev Oncol Hematol 2013;87:146-71.
2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-11.
3. Rosenwald A, Staudt LM. Clinical translation of gene expression profiling in lymphomas and leukemias. Semin Oncol 2002;29:258-63.
4. Visco C, Li Y, Xu-Monette ZY, Miranda RN, Green TM, Li Y, *et al.* Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: A report from the international DLBCL rituximab-CHOP consortium program study. Leukemia 2012;26:2103-13.
5. Meyer PN, Fu K, Greiner TC, Smith LM, Delabie J, Gascoyne RD, *et al.*

Immunohistochemical methods for predicting cell of origin and survival in patients with diffuse large B-cell lymphoma treated with rituximab. J Clin Oncol 2011;29:200-7.
6. Hans CP, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, Ott G, *et al.* Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. Blood 2004;103:275-82.
7. Castillo JJ, Beltran BE, Song MK, Ilic I, Leppa S, Nurmi H, *et al.* The hans algorithm is not prognostic in patients with diffuse large B-cell lymphoma treated with R-CHOP. Leuk Res 2012;36:413-7.
8. Ott G, Ziepert M, Klapper W, Horn H, Szczepanowski M, Bernd HW, *et al.* Immunoblastic morphology but not the immunohistochemical GCB/nonGCB classifier predicts outcome in diffuse large B-cell lymphoma in the RICOVER-60 trial of the DSHNHL. Blood 2010;116:4916-25.
9. Choi WW, Weisenburger DD, Greiner TC, Piris MA, Banham AH, Delabie J, *et al.* A new immunostain algorithm classifies diffuse large B-cell lymphoma into molecular subtypes with high accuracy. Clin Cancer Res 2009;15:5494-502.
10. Natkunam Y, Farinha P, Hsi ED, Hans CP, Tibshirani R, Sehn LH, *et al.* LMO2 protein expression predicts survival in patients with diffuse large B-cell lymphoma treated with anthracycline-based chemotherapy with and without rituximab. J Clin Oncol 2008;26:447-54.
11. de Jong D, Rosenwald A, Chhanabhai M, Gaulard P, Klapper W, Lee A, *et al.* Immunohistochemical prognostic markers in diffuse large B-cell lymphoma: Validation of tissue microarray as a prerequisite for broad clinical applications – A study from the lunenburg lymphoma biomarker consortium. J Clin Oncol 2007;25:805-12.
12. Quinlan R. The Book C4. 5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.
13. PCViz: Pathway Commons Network Visualizer. Memorial Sloan-Kettering Cancer Center and the University of Toronto; 2016. Available from: http://www.pathwaycommons.org/pcviz/#neighborhood/FOXP1, MUM1. [Last accessed on 2016 Jun 20].
14. Gutiérrez-García G, Cardesa-Salzmann T, Climent F, González-Barca E, Mercadal S, Mate JL, *et al.* Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large B-cell lymphoma treated with immunochemotherapy. Blood 2011;117:4836-43.
15. Allman D, Jain A, Dent A, Maile RR, Selvaggi T, Kehry MR, *et al.* BCL-6 expression during B-cell activation. Blood 1996;87:5257-68.
16. Lossos IS, Jones CD, Warnke R, Natkunam Y, Kaizer H, Zehnder JL, *et al.* Expression of a single gene, BCL-6, strongly predicts survival in patients with diffuse large B-cell lymphoma. Blood 2001;98:945-51.
17. Maeshima AM, Taniguchi H, Fukuhara S, Morikawa N, Munakata W, Maruyama D, *et al.* Bcl-2, Bcl-6, and the international prognostic index are prognostic indicators in patients with diffuse large B-cell lymphoma treated with rituximab-containing chemotherapy. Cancer Sci 2012;103:1898-904.
18. Coutinho R, Clear AJ, Owen A, Wilson A, Matthews J, Lee A, *et al.* Poor concordance among nine immunohistochemistry classifiers of cell-of-origin for diffuse large B-cell lymphoma: Implications for therapeutic strategies. Clin Cancer Res 2013;19:6686-95.