Human Cancer Biology

Clinical Cancer Research

Poor Concordance among Nine Immunohistochemistry Classifiers of Cell-of-Origin for Diffuse Large B-Cell Lymphoma: Implications for Therapeutic Strategies

Rita Coutinho[1], Andrew James Clear[1], Andrew Owen[1,2], Andrew Wilson[1], Janet Matthews[1], Abigail Lee[1,2], Rute Alvarez[3], Maria Gomes da Silva[3], José Cabeçadas[4], Maria Calaminici[1,2], and John G. Gribben[1]

Abstract

Purpose: The opportunity to improve therapeutic choices on the basis of molecular features of the tumor cells is on the horizon in diffuse large B-cell lymphoma (DLBCL). Agents such as bortezomib exhibit selective activity against the poor outcome activated B-cell type (ABC) DLBCL. In order for targeted therapies to succeed in this disease, robust strategies that segregate patients into molecular groups with high reliability are needed. Although molecular studies are considered gold standard, several immunohistochemistry (IHC) algorithms have been published that claim to be able to stratify patients according to their cell-of-origin and to be relevant for patient outcome. However, results are poorly reproducible by independent groups.

Experimental Design: We investigated nine IHC algorithms for molecular classification in a dataset of DLBCL diagnostic biopsies, incorporating immunostaining for CD10, BCL6, BCL2, MUM1, FOXP1, GCET1, and LMO2. IHC profiles were assessed and agreed among three expert observers. A consensus matrix based on all scoring combinations and the number of subjects for each combination allowed us to assess reliability. The survival impact of individual markers and classifiers was evaluated using Kaplan–Meier curves and the log-rank test.

Results: The concordance in patient's classification across the different algorithms was low. Only 4% of the tumors have been classified as germinal center B-cell type (GCB) and 21% as ABC/non-GCB by all methods. None of the algorithms provided prognostic information in the R-CHOP (rituximab plus cyclophosphamide–adriamycin–vincristine–prednisone)–treated cohort.

Conclusion: Further work is required to standardize IHC algorithms for DLBCL cell-of-origin classification for these to be considered reliable alternatives to molecular-based methods to be used for clinical decisions. *Clin Cancer Res; 19(24); 6686–95. ©2013 AACR.*

Introduction

Diffuse large B-cell lymphoma (DLBCL) represents a heterogeneous group of lymphoid malignancies with distinct oncogenic events and clinical behavior that cannot be unraveled by morphology and immunophenotype (1–4). This biologic segregation helps to explain the heterogeneous responses to standard treatment and provides a rationale for investigation of novel targeted therapies. Emerging data support the notion that the two main DLBCL molecular groups, the activated B-cell like (ABC) and the germinal center B-cell like (GCB) DLBCLs, benefit from different treatment approaches, with agents including bortezomib (5), lenalidomide (6, 7), or ibrutinib (8) seeming particularly active against the worse-prognosis ABC subtype.

Although gene expression profiling (GEP) is the widely used and accepted method for molecular stratification of DLBCL at the bench, this technique has only recently been incorporated into clinical trials for treatment stratification. The *REMoDLB* phase III clinical trial (NCT01324596) aims at determining whether the addition of bortezomib to standard R-CHOP (rituximab plus cyclophosphamide–adriamycin–vincristine–prednisone) improves event-free survival (EFS) and whether that benefit is related to the molecular features of the tumor cells, which is being characterized by GEP in the formalin-fixed paraffin-embedded (FFPE) tissue. However, because its application is restricted to research purposes, there is presently a lack of

Authors' Affiliations: [1]Department of Hemato-Oncology, Barts Cancer Institute, Queen Mary University of London; [2]Department of Histopathology, Barts Health NHS Trust, Royal London Hospital, London, United Kingdom; Departments of [3]Hematology, and [4]Pathology, Portuguese Institute of Oncology, Lisbon, Portugal

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: John G. Gribben, Barts Cancer Institute, Queen Mary University of London, John Vane Science Centre, Charterhouse Square, London EC1M 6BQ, United Kingdom. Phone: 44-20-7882-3805; Fax: 44-20-7882-3891; Email: j.gribben@qmul.ac.uk

doi: 10.1158/1078-0432.CCR-13-1482

Translational Relevance

Molecular characterization is opening opportunities for personalized therapy in poor risk diffuse large B-cell lymphoma (DLBCL). Clinical trials using gene expression profiling (GEP) as stratifiers are under way. Immunohistochemistry (IHC) is attractive as a surrogate for molecular stratification in DLBCL and the Hans algorithm is being used to define DLBCL of the activated B-cell type (ABC) in clinical trials offering NF-κB targeting agents. However, the applicability of IHC classifiers has been questioned. We investigated nine IHC algorithms in a large dataset of diagnostic DLBCLs and report a high degree of disagreement in classifying a single patient by all methods. Moreover, none of the methods was able to identify different prognostic groups within R-CHOP (rituximab plus cyclophosphamide–adriamycin–vincristine–prednisone)–treated subjects. We suggest that the application of IHC as an alternative to molecular-based approaches should be used with caution. Collaborative IHC studies to provide procedural guidelines for use in the clinical arena are warranted.

standardized methodology for GEP analysis, which can lead to variable results both at the inter- and intralaboratory levels. This issue, which may impact GEP results and patient care, is generally unreported.

The absence of a routine methodology for GEP-based cell-of-origin assessment has encouraged investigators to develop immunohistochemistry (IHC)-based approaches for the molecular classification in DLBCL. In 2005, Hans and colleagues (9) established the first IHC algorithm, with supposed high sensitivity for GEP classification. Subsequently, eight further strategies (refs. 10–15; Fig. 1) have been published, all of which reported a better concordance with molecular-based classification and an ability to segregate two groups with different outcome. However, many investigators continue to question their clinical applicability (16–22). We provide an up-to-date systematic comparison of nine IHC scores for molecular classification in a new large dataset of diagnostic DLBCL. Our primary aim was to test the reliability of these methodologies in individual cases in this cohort. We report that none of the nine algorithms used is able to predict outcome in this representative dataset of R-CHOP patients. Although this has been described using some of the classifiers (18, 19, 22, 23), this is the first study demonstrating it for all methods developed to date. Moreover, the concordance in classifying a single tumor into GCB and non-GCB/ABC across all algorithms was statistically very low.

Materials and Methods

Patient characteristics

Ethical approval for this study was obtained from the local regional ethics boards. Patient selection was dependent on the availability of good-quality FFPE tissue of the

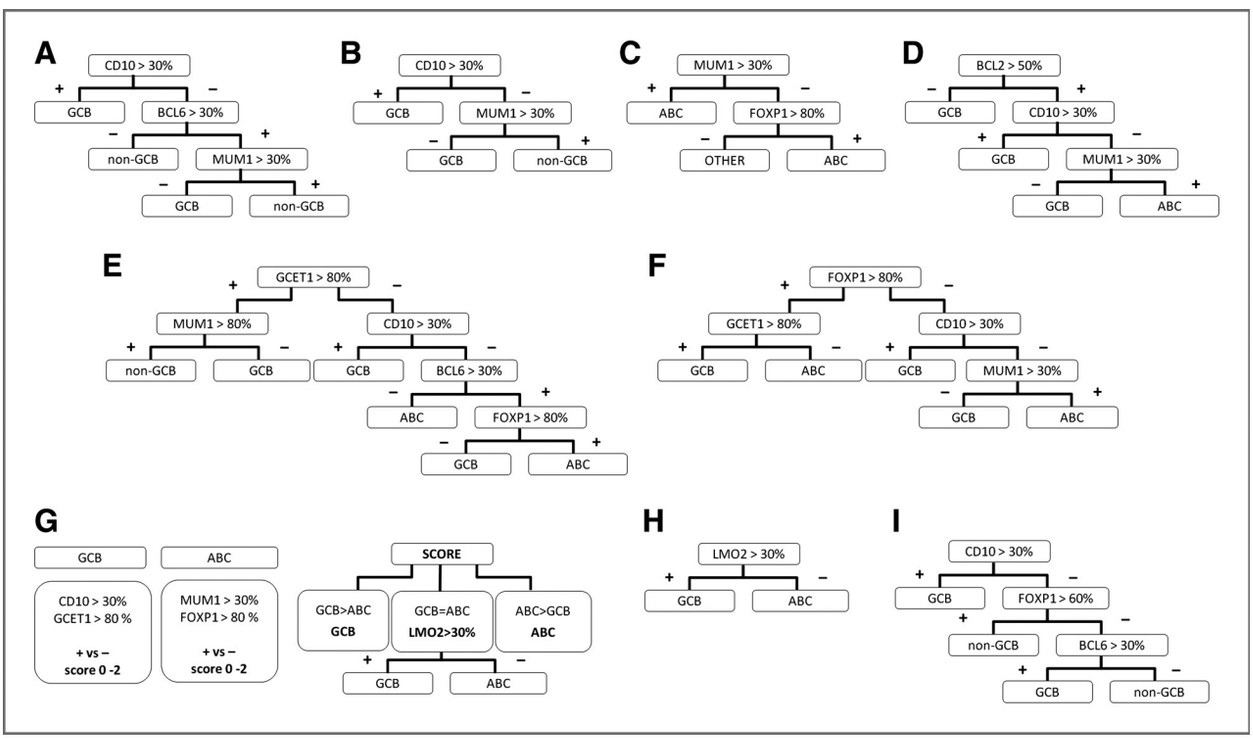

Figure 1. Algorithms applied in the current study. A, Hans; (B) Hans modified; (C) Nyman; (D) Muris; (E) Choi; (F) Choi modified; (G) Tally; (H) Natkunam; and (I) Visco–Young.

diagnostic biopsy and clinical and follow-up data. Only cases of *de novo* DLBCL were included. Patients with an immunodeficiency-associated lymphoma, central nervous system, or primary mediastinal lymphomas were excluded from the study.

From 651 patients with *de novo* DLBCL diagnosed at St. Bartholomew's Hospital (London, United Kingdom) between 1977 and 2009, we identified 218 patients with available diagnostic biopsy material amenable for array. Seventy-one of these patients were treated in the rituximab era with R-CHOP and had extended clinical and follow-up data. The remaining patients were treated with different approaches, from anthracycline-based therapy to palliative care, and for this reason were excluded from outcome analysis. To enhance the R-CHOP–treated cohort, further 80 patients from the Portuguese Institute of Oncology (Lisbon, Portugal) who had high-quality diagnostic biopsy material were included in this study. Therefore, included in the study were samples from 151 R-CHOP chemo-immunotherapy–treated patients and 147 patients treated with chemotherapy alone.

Clinical data, including response to chemotherapy and follow-up time, are detailed in Table 1 only for the R-CHOP–treated cohort, for which outcome analysis was performed. No significant differences were observed about main initial features and outcome between patients with available tissue suitable for inclusion in the tissue microarrays (TMA) and the remainder (data not shown).

Table 1. Clinical features of the R-CHOP series

Characteristics	*n* (%)
Age	
Median (range)	62 (16–86)
>60 y	81 (55)
Male sex	77 (52)
B-symptoms	38 (26)
≥2 Extranodal sites	24 (16)
Ann Arbor stages III and IV	81 (55)
ECOG performance status ≥2	23 (16)
High LDH	97 (66)[a]
IPI	54 (37)/33 (23)
(Low/low-int/high-int/high)	37 (26)/20 (14)
Treatment response	
CR	110 (75)
PR	22 (15)
SD/PD	15 (10)
Mortality	47 (32)
Follow-up (months, range)	53 (4–71)

NOTE: Clinical characteristics were available for 147 patients. Significant differences between the two R-CHOP groups are detailed in the results.

Abbreviations: ECOG, Eastern Cooperative Oncology Group; high-int, high-intermediate; low-int, low-intermediate; *n*, number; PD, progressive disease; PR, partial response; SD, stable disease.

[a]A significantly higher number of patients diagnosed at IPO had high LDH at diagnosis by the Fisher exact test.

TMA and IHC

TMAs were prepared from the paraffin blocks using a manual tissue arrayer (Beecher Scientific) in both institutions. Triplicate or duplicate 1 to 1.5 mm^2 cores were taken from regions of biopsy material rich in malignant cells identified on hematoxylin and eosin (H&E)–stained sections. Tonsils were cored in all TMAs. Staining for the CD20 was performed to confirm adequate tumor representation.

After dewaxing, blocking in hydrogen peroxide/methanol solution, rehydration, and antigen retrieval, the slides were subjected to immunostaining. Primary antibody reaction for CD10, BCL6, BCL2, MUM1, FOXP1, GCET1, and LMO2 was detected using a peroxidase-labeled system (Super-Sensitive Polymer-HRP IHC Detection System; BioGenex). An immunologic amplification method (CSA II, Catalyzed Amplification System; Dako) was used exclusively for BCL6.

IHC for GCET1, FOXP1, and LMO2 required optimization. Appropriate controls for titration and antigen retrieval are provided by the manufacturer. Serial 1/500, 1/250, and 1/100 antibody dilutions and three antigen retrieval techniques were tested.

In each batch of staining, tonsil sections were analyzed simultaneously for all markers. All IHC studies were performed in the same laboratory.

Primary antibodies and conditions of use are provided in Supplementary Table S1. Representative negative, positive, and control cases are provided in Supplementary Fig. S1.

Scoring and analysis

Slides were scanned using the Hamamatsu Virtual Slide Scanner NanoZoomer 2.0 and viewed using the NDP.scan software. Cores were visualized on a computer screen at low and high power in an initial joint training session (R. Coutinho, A. Lee, and M. Calaminici) in order for scoring criteria to be standardized for each marker. Each case was scored as positive or negative according to the cutoff points defined in the original publications and detailed in Fig. 1. It is important to stress that this means that some antibodies were analyzed for more than one cutoff point, as described in Table 2. As recommended by the Lunenburg Lymphoma Biomarker Consortium guidelines (24), negative cases with absent internal controls were considered unclassifiable. This and the absence of whole cores in the TMA were the primary causes for the inability to score (unclassifiable cases are detailed for each antibody in Table 2). Whenever individual cores of a given case showed nonconcordant results, the core with highest large cell infiltration was used. After each antibody was assessed by all observers, a discussion meeting was organized to reach consensus on discordant cases. In almost all cases a 3:0 decision was reached, but in less than 5% of the cases a 2:1 score was accepted, particularly taking into account the opinion of the most

Table 2. Single marker analysis

Antigens	Positive [*n* (%)]	Negative [*n* (%)]	Unclassifiable (*n*)
CD10 (30%)	41 (28)	106 (72)	4
BCL6 (30%)	86 (61)	54 (39)	11
MUM1 (30%)	109 (76)	35 (24)	7
MUM1 (80%)	60 (42)	84 (58)	7
GCET1 (30%)[a]	29 (20)	118 (80)	4
GCET1 (80%)	12 (8)	135 (92)	4
FOXP1 (30%)	116 (82)	25 (18)	10
FOXP1 (60%)	102 (72)	39 (28)	10
FOXP1 (80%)[b]	77 (55)	64 (45)	10
LMO2 (30%)	79 (56)	62 (44)	10
BCL2 (50%)	96 (69)	43 (31)	12

NOTE: Absolute number and percentage of classifiable cases are detailed for each antigen. Staining and analysis were performed centrally.

Abbreviation: *n*, number.

[a]A significantly higher proportion of GCET1$^+$ cases (30% cutoff) was detected in patients diagnosed at Bart's Hospital ($P = 0.006$).

[b]A significantly higher number of patients diagnosed in IPO were considered positive for FOXP1 at a cutoff of 80% ($P = 0.001$).

experienced hematopathologist (M. Calaminici). Inter-rater agreement was higher than 95% in all antibodies except BCL6 (92%). As expected, consensus was higher for the remaining antibodies in use in the diagnostic setting, such as CD10, BCL2, and MUM1. LMO2 shows a nuclear distribution and is also expressed by T cells so analysis also included nuclear size. FOXP1 scoring was difficult in some cases due to the background staining and inter-patient differences in staining intensity.

Statistical analysis

To render outcome analysis more relevant for the clinical community, all results about single marker expression and algorithm distribution refer only to the more recently diagnosed and R-CHOP–treated cohort. Differences in clinical characteristics between the two R-CHOP series were tested using a χ^2 or Fisher exact test, when appropriate. A variable number of cases failed to yield reliable staining results for the reasons already detailed and were excluded from the analysis.

All cases included in the arrays (298 diagnostic samples) were used in the serial comparison of algorithm performance. A consensus matrix based on the nine classifiers and number of subjects for each combination was built. Concordance across all methods was measured using the general κ statistics.

In the univariate analysis, log-rank tests were performed. All parameters of the International Prognostic Index (IPI) were assessed for prognostic impact. Relevant interactions of the markers studied with the IPI factors were assessed. Cox logistic regression models included all variables with a $P \geq 0.2$ on univariate analysis.

The outcomes, measured from the date of diagnosis to the occurrence of event or date of last follow-up, were overall survival (OS), the event being death from any cause and EFS, the event being failure of treatment [including not achieving complete remission (CR) or death of any cause]. Median follow-up was calculated for patients alive at last follow-up. Statistical analysis was performed using SPSS version 19.0 (SPSS) and Prism version 5.03 (GraphPad Software).

Results

Analysis of individual markers

Ever since the GEP-based cell-of-origin classification was described, a growing number of immunohistochemical approaches have been used to attempt to define GCB and ABC subtypes. It is generally accepted that no single protein expression seems to be able to mirror the GEP classification. Results from single protein expression (positive, negative, and unclassifiable cases) in our series are illustrated in Table 2.

CD10 expression was detected in 41 (28%) of our patients, in keeping with previous results (9, 10, 13, 17, 19, 25). CD10 expression enriched for younger patients ($P = 0.03$), whereas negative cases had higher lactate dehydrogenase (LDH; $P = 0.02$). However, no differences in the IPI distribution were detected. We observed a positive correlation between CD10 expression and other GCB-"specific" proteins, including BCL6 (Pearson $r = 0.311$; $P < 0.001$) and GCET1 (Pearson $r = 0.45$; $P < 0.001$), and a negative correlation with the post-GC marker MUM1 (Pearson's $r = -0.164$; $P = 0.05$).

Both GCB and ABC DLBCLs can harbor genetic aberrations involving *BCL6*, leading to protein overexpression. Nevertheless, BCL6 is considered a GC marker. BCL6 expression was comparable with previous reports (9, 19, 25, 26). A positive correlation was observed with other GCB-related proteins (CD10; GCET1, Pearson $r = 0.39$, $P < 0{,}001$; and LMO2, Pearson $r = 0.51$, $P < 0.001$) and the ABC marker FOXP1 (Pearson $r = 0.194$; $P = 0.031$). BCL6$^+$ patients were more likely to have low-risk IPI (log-rank $P = 0.05$), and in this IPI subgroup BCL6 expression alone conferred a better 3-year OS (79% vs. 93%; $P = 0.04$). When the whole dataset was analyzed, BCL6 was unable to differentiate patients with distinct outcome.

Expression of the post-GC MUM1 at a cutoff of 30% was detected in 109 (76%) of our subjects, higher than previously reported (9, 23). However, using an 80% cutoff, 42% of the patients were deemed positive for MUM1, which is in keeping with the Choi data (13). Moreover, this higher cutoff improved correlation with other post-GC markers, such as FOXP1 (Pearson $r = 0.19$–0.37; $P < 0.01$). A significant correlation between BCL2 and MUM1 expression was also documented (Pearson $r = 0.36$–0.37; $P < 0.01$).

GCET1 protein expression is restricted to a subset of GCB cells (27) and should specifically identify GCB DLBCLs.

However, expression of this marker was detected in only 8% of cases when assessed at a cutoff point of 80%, and this only increased to 20% at the lower cutoff of 30%. At both cutoff points, although positive patients were more likely to be younger ($P = 0.003$), there was no association with IPI. At the 30% cutoff for expression, positive correlations with other GC markers, including CD10 and LMO2 (Pearson $r = 0.34$; $P < 0.001$), and negative correlation with MUM1 (Pearson $r = -0.19$; $P = 0.03$) were detected, with weaker correlations at the higher cutoff point.

FOXP1 has been detected at high levels in cases lacking GCB markers and expressing BCL6 and MUM1 (28). In our cohort, 82% of the patients had more than 30% and 72% had more than 60% of FOXP1$^+$ cells. At a higher cutoff of 80%, we noted a higher proportion of FOXP1$^+$ Portuguese patients (67% vs. 40%; $P = 0.001$).

LMO2 expression is restricted to the nucleus of normal GCB cells (29) in lymph nodes and to a subset of GCB-DLBCLs. Natkunam and colleagues (11) proposed that LMO2 alone has a high predictive power for GCB/non-GCB allocation. We detected LMO2 staining in 79 (56%) cases, which is in keeping with the original data. Significant correlations with other GC molecules, particularly BCL6 and GCET1, were observed.

BCL2 is commonly targeted in DLBCL, with half of GCB and the majority of ABC-DLBCLs having *BCL2* overexpression (4). Protein expression has been associated with poor outcome in most studies (21, 30–32). However, its survival impact seems to be modulated by rituximab (33, 34). In our series, BCL2 was expressed in 96 (69%) patients, with no documented associations with clinical factors or place of diagnosis.

Algorithm classification: distribution and consistency

According to the original and most widely used Hans method (9), we classified 53 (38%) of our subjects as GCB. CD10 expression was determinant for this allocation, as 41 of these patients were classified as positive for this antigen. This is in keeping with the original publication. The remaining 12 cases were all BCL6$^+$ and MUM1$^-$. Within the non-GCB cohort, BCL6 expression was detected in half of the cases. The Hans allocation was similar between the two R-CHOP cohorts and no association with clinical characteristics was found.

The modified Hans was proposed to decrease inconsistency brought about by the anti-BCL6 antibody (14). According to modified Hans criteria, 61 (42%) patients were classified as GCB, including all patients with GCB from the Hans method and a further 8 patients scored CD10$^-$/MUM1$^-$.

The Choi classifier (13) relies on the expression of three antigens (GCET1, MUM1, and CD10) for initial allocation and a further two (BCL6 and FOXP1) for final decision. Cutoff points were adopted according to appropriate detection of single-series molecularly profiled patients (13). Using the cutoff points defined in their article, allocation into GCB and non-GCB groups in our cohort was 65 (45%) and 79 (55%) patients, respectively.

With the intention of simplifying this complex algorithm, Meyer and colleagues (14) proposed the modified Choi (Fig. 1), which allocated 35 (25%) and 105 of our patients to GCB and non-GCB groups, respectively. Compared with the Choi algorithm, the modified version reclassified both GCB (64 cases) and non-GCB cases (8 cases). Because of the different proportion of FOXP1$^+$ patients between the two R-CHOP series (see above), we observed a higher proportion of ABC cases in the Portuguese series, using the Choi ($P = 0.05$) and the modified Choi ($P = 0.01$) criteria (see also Supplementary Table S2). The Choi ABC cohort was enriched for older patients ($P = 0.01$).

From the clinical point of view, it looks more relevant to use a method that is highly sensitive at identifying patients with ABC, the subgroup that may be more amenable to targeting with new agents. Nyman and colleagues (12) proposed a method with this purpose, using only post-GC antibodies (MUM1 and FOXP1). In our series, only 18 patients (13%) have been considered non-ABC using this approach. No associations with clinical characteristics were detected.

The Muris classifier (10) is the only that uses BCL2. In our series, 71 (50%) patients were classified as GCB, 43 of which were BCL2$^-$, suggesting that this marker plays a predominant role at defining this cell-of-origin subgroup.

The Tally algorithm (14) has the unique feature of attributing similar weight to GCB- and post-GCB–specific markers for allocation. There was a predominance of cases (110, 78%) that were classified as ABC. In 34 cases, LMO2 expression was used for decision, as it was positive in half of the cases.

Finally, Visco and colleagues (15) recently launched another method with an increased overlap with GEP data. Similar to Hans, CD10 plays the central role for GCB allocation (41 of 53 patients with GCB were CD10$^+$). About non-GCB allocation, BCL6 plays less of a role than in the Hans classifier as in only 15 CD10$^-$/FOXP1$^-$ patients was BCL6 expression taken into account for allocation. As with the Choi classifier, there was an enrichment for older patients in the Visco ABC subset ($P = 0.02$).

To address the question, which has been so far left unanswered by many groups, of how every individual patient is classified across all methods, we then performed a parallel classification of all tumors using the nine cell-of-origin algorithms. Results for all classifiers were available for 242 of the 298 cases (81%). Surprisingly, only 4.1% of the tumors were classified as GCB by all methods. The degree of agreement in allocation of patients to the non-GCB group was significantly higher, with 21% of patients being allocated to this group by all methods and 20.6% being classified as non-GCB by all methods except one—either the Choi (2 cases), the modified Choi (2 cases), the Natkunam (33 cases), or the Muris (13 cases) algorithms. Of note, the last two are the only methods in which allocation to the GCB subset was higher.

We then sought to assess pairwise agreement using the general κ statistics, a method that tests for interscoring reliability. The κ is considered a robust measure as it takes

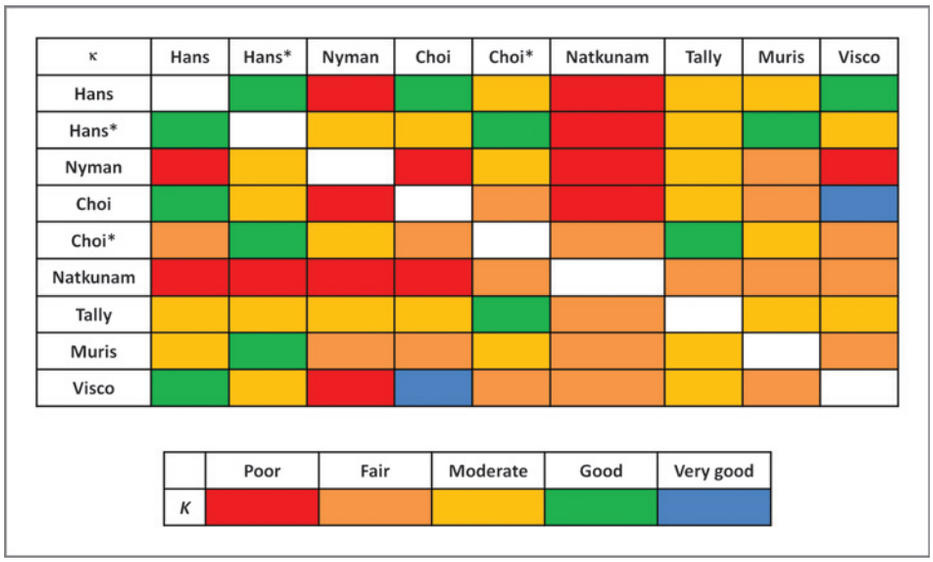

Figure 2. Pairwise agreement according to κ statistics. *, Modified.

into account the agreement occurring by chance. Figure 2 illustrates the strength of agreement among all scoring systems. Poor and fair κ values were detected in 44.4% on pairwise concordance assessment; and in only 20% was κ good or very good. The Natkunam algorithm is the least concordant with the remaining, showing a poor agreement with four algorithms and only fair agreement with the other four. The highest level of agreement was found between the Choi and the Visco algorithms ($\kappa = 0.85$). From all the methods investigated, the Hans and the Hans modified methods exhibited the highest degree of consistency with other algorithms.

Survival analysis

For the purpose of outcome analysis, we considered that only patients treated with the current standard of care should be included and, therefore, this analysis was performed on the 151 samples from R-CHOP–treated patients. As the IPI remains the most robust prognostic discriminator in DLBCL, we assessed whether its individual variables or subgroups had a role in predicting outcome in the R-CHOP cohort (Table 4). On univariate analysis, age, stage, performance status, and IPI groups were significant in predicting OS, whereas number of extranodal sites, staging, performance status, and IPI groups were significant in predicting EFS. No immunohistochemical marker alone achieved significance for outcome prediction in R-CHOP–treated patients. Although patients expressing FOXP1 (60% cutoff point) had a lower OS (72% vs. 82%; $P = 0.09$) and patients expressing BCL2 had a lower EFS (57% vs. 77%; $P = 0.06$), none reached significance on forward stepwise multivariate analysis together with either the IPI factors or the IPI subgroups.

As described in Table 3, in our experience and corroborating others data, none of the algorithms rival the IPI for OS or EFS prediction in R-CHOP–treated patients. We have also looked at survival differences between patients classified as either GCB or non-GCB by all methods versus the remaining patients with heterogeneous classification. Although OS was similar among groups, 3-year EFS was significantly ($P = 0.004$) better for the GCB set (100%) compared with the ABC set (78%) or the remaining patients (60%).

As survival was similar across all classifiers, we sought to determine whether outcome stratification could be improved by analyzing the expression of additional proteins not included in the original algorithms. If this is demonstrated, it would suggest that IHC classifiers are oversimplified methods for the purpose of outcome stratification. As an example, BCL2 expression was associated with worse EFS in "GCB" cases only, when incorporated into the Hans (54% vs. 88%; $P = 0.006$), Hans modified (52% vs. 83%; $P = 0.009$), Visco (54% vs. 86%; $P = 0.01$), Natkunam (45% vs. 78%; $P = 0.02$), and Choi modified (53% vs. 86%; $P = 0.02$) methods. Similarly, expression of the post-GC markers FOXP1 and MUM1 was associated with worse survival in those cases defined as GCB using the Hans and Natkunam algorithms (data not shown).

Discussion

Clinical trials are under way investigating the use of novel agents in subsets of patients with DLBCL, particularly in the poor risk ABC subtype. It is well recognized that GEP is the standard method to designate patients into molecular subsets and clinical trials using GEP for this purpose will clarify the utility of targeted therapies in the clinical setting. However, the applicability of molecular classification into clinical practice will require a robust, affordable, and reproducible methodology for designation. It was hoped that immunohistochemical approaches would be useful surrogates for the classification of DLBCL subsets, would be readily applicable in clinical practice, and would be incorporated into diagnostic workup within hematopathology clinical laboratories. However, on the basis of previous work and the data presented here, we suggest that much work needs to be done to standardize the Hans and other

Table 3. Distribution of R-CHOP–treated patients according to the nine IHC classifiers

Classifiers	GCB [*n* (%)]	Non-GCB/ABC [*n* (%)]	Unclassifiable (*n*)	*P*
Hans	53 (38)	87 (62)	11	NS
Hans modified	61 (42)	84 (58)	6	NS
Choi	65 (45)	79 (55)	7	0.05
Choi modified	35 (25)	105 (75)	11	0.01
Natkunam	79 (56)	62 (44)	10	NS
Nyman	18 (13)	126 (87)	7	0.03
Muris	71 (50)	70 (50)	10	0.03
Tally	30 (21)	110 (78)	11	NS
Visco–Young	53 (37)	90 (63)	8	NS

NOTE: Unclassifiable cases were excluded for percentage calculation. Significant differences observed in the cohorts from the two institutions according to the Fisher exact test are detailed in the text.
Abbreviations: *n*, number; NS, not significant.

IHC methods that currently should be considered unreliable surrogates for molecular classification in DLBCL.

Here, we systematically analyzed the nine IHC DLBCL classification algorithms in a representative dataset of diagnostic DLBCL and our objective was to describe how each individual case would be scored by all classifiers. Although this study would have been enhanced by the availability of GEP as a "gold-standard," the methodology used here does not require such a comparison, as we sought to examine the robustness of the more commonly used IHC algorithms and their ability to classify DLBCL compared with each other. Classifier distribution in our cohort was heterogeneous. Using the Hans method, we report a GCB subtype in only 38% of patients, in line with the findings of other authors (18, 19). CD10 plays a critical role for GCB allocation in the Hans method, and we believe a proportion of molecularly defined GCB cases are being allocated as non-GCB due to higher expression of MUM1. Using the complex Choi classifier, both GCET1 and MUM1 had almost no impact on allocation, making this method in our experience very similar to that of Visco. In our series, only 18 patients were considered non-ABC using the Nyman method, suggesting too low specificity for ABC cases. Similarly, a predominant number of cases (110) were classified as ABC using the Tally

Table 4. Survival analysis according to clinical characteristics, IPI, and cell-of-origin IHC classifiers

	3-y OS		3-y EFS	
Variables	**%**	***P***	**%**	***P***
Sex (male vs. female)	74 vs. 77	NS	58 vs. 72	NS
Age (<60 vs. >60 y)	83 vs. 68	0.03	66 vs. 64	NS
Number extranodal sites (<2 vs. ≥2)	79 vs. 53	0.09	71 vs. 33	0.003
Ann Arbor stage (I–II vs. III–IV)	87 vs. 63	0.001	83 vs. 49	<0.001
ECOG performance status (<2 vs. ≥2)	81 vs. 48	<0.001	73 vs. 23	<0.001
LDH (low vs. high)	79 vs. 72	NS	73 vs. 60	NS
IPI (low/low-int/high-int/high)	90 vs. 72 vs. 73 vs. 45	<0.001	79 vs. 80 vs. 53 vs. 21	<0.001
Algorithms (GCB vs. non-GCB/ABC)				
Hans	77 vs. 74	NS	66 vs. 66	NS
Hans modified	75 vs. 75	NS	63 vs. 66	NS
Choi	75 vs. 75	NS	63 vs. 62	NS
Choi modified	74 vs. 75	NS	63 vs. 62	NS
Muris	79 vs. 71	NS	67 vs. 64	NS
Nyman	76 vs. 75	NS	75 vs. 64	NS
Tally	78 vs. 73	NS	72 vs. 64	NS
Natkunam	71 vs. 78	NS	59 vs. 74	NS
Visco–Young	75 vs. 75	NS	62 vs. 68	NS

Abbreviations: ECOG, Eastern Cooperative Oncology Group; High-int, high-intermediate; Low-int, low-intermediate; NS, not significant.

method, driven by the expression pattern of GCET1 and MUM1 in our series.

Taken all classifiers together, we document an extremely low concordance across all techniques, especially for those more likely to represent the GCB subtype. This serial comparison of algorithm performance was done in all cases included in our arrays. This analysis does not imply any comparison across samples from different tumors (with inherent differences in the quality of the material and consequently in the results obtained), but only how each method performs within the same tumor sample for reaching the same endpoint.

Moreover, results were made more robust by using the κ statistics, which takes into account the agreement occurring by chance. Scoring allocation seemed more consistent across all methods for the non-GCB group. Previous analyses using paired GEP and IHC also demonstrated that the proportion of misclassified cases by IHC compared with GEP was higher when defining the GCB subtype (14, 19). A previous report has suggested a good concordance between the Choi and Hans algorithms and GEP (14). However, this study developed a new algorithm (Tally), which had even better concordance with GEP. Despite these findings, we demonstrate low concordance between the Hans and Tally algorithms.

We noted a correlation of expression of GCB markers, including CD10, BCL6, and LMO2. However, BCL6 expression was associated with both GCB and ABC markers, suggesting that expression of this protein is not entirely restricted to GCB or ABC cells. BCL6 detection requires amplification, making it even more difficult to standardize among laboratories (24). We identified a smaller proportion of cases expressing the remaining GCB-identifying protein GCET1 than previously reported. In the original study (13), an amplification method was used to enhance GCET1 staining, whereas others used a different antigen retrieval strategy (14). This, together with the staining pattern of GCET1, might help to explain our results. However, as this antigen has been studied by relatively few groups, we propose that more experience has to be gathered on patterns of expression and optimal staining procedures for GCET1 before this is incorporated widely into DLBCL classification.

Although IHC assessment for the post-GC marker MUM1 is highly sensitive to laboratory variations and inter-interpreter scoring, its expression is incorporated in many algorithms. We noted more than 30% expression of MUM1 in a higher proportion of patients than previous studies and this cannot be explained by commonly reported reasons such as nonspecific cytoplasmic background staining and target cell artifacts (24). Choi and colleagues (13) claimed that a higher cutoff level of expression of 80% was required to achieve high specificity for ABC cases. Using this cutoff point, 42% of our patients were deemed positive for MUM1 expression, in keeping with the Choi data. This, however, highlights the difficulties of standardizing results based on arbitrary cutoffs. We also documented significant differences in the proportion of cases expressing FOXP1 in 80% of malignant B cells between the Portuguese and the English datasets. Although the biopsies were obtained in the two countries, staining and analysis were performed in the same laboratory. Whether this is a reflection of different fixation and storing methods, use of arbitrary cutoffs or a true ethnic difference still has to be demonstrated and further population studies will be required to address this.

As has already been demonstrated in other cohorts (35), we detected associations between clinical factors and protein expression. In both the Choi and Visco classifiers, the ABC subset was enriched for older patients. This suggests that clinical prognostic factors might interact with biologic predictors such as the cell-of-origin classification for DLBCL.

The prognostic impact of IHC classifiers has been questioned by many authors. Nevertheless, our secondary aim was to perform survival analysis based on these algorithms. It is recognized that samples collected over a long period of time have differences in quality that might impact adequate interpretation of immunostaining results. This is particularly important when survival analysis is undertaken. Having this in mind we analyzed only those cases treated with R-CHOP, as this is the only clinically relevant treatment in current practice. None of the IHC classifiers was able to predict outcome in this series. This has been demonstrated by others, particularly using the Hans method (16–19, 21, 23, 35). We acknowledge that the R-CHOP series analyzed (151 cases) is limited and the number of events registered during the study period renders it underpowered to detect differences in survival between the two groups. However, we believe this fact supports our hypothesis that IHC classifiers are inadequate to recognize the molecularly defined DLBCLs. Analyzing individual markers not included in the algorithms can improve outcome prediction, as has been demonstrated by us and others (36) using BCL2.

Taking our data into consideration, it is difficult to recommend any specific classifier for further use in the clinical practice. Although many potential reasons for inconsistent results using IHC have been reported (24, 37), no definitive procedural consensus is available for implementation in clinical practice. Our study challenges the use of any IHC classifier for subclassification in DLBCL, and it is important to resolve this, as the cell of origin in DLBCL provides not only prognostic information but also offers the window for targeted therapies to improve outcome in this disease.

Disclosure of Potential Conflicts of Interest

J.G. Gribben has honoraria from Speakers Bureau of Roche, Celgene, and Pharmacyclics and is a consultant/advisory board member of Celgene. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: R. Coutinho, M. Calaminici, J.G. Gribben
Development of methodology: R. Coutinho, A.J. Clear, J.G. Gribben
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): R. Coutinho, A.J. Clear, A. Owen, J. Matthews, A. Lee, R. Alvarez, M. Gomes da Silva, J. Cabeçadas, M. Calaminici, J.G. Gribben
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): R. Coutinho, A. Wilson, J. Matthews, A. Lee, M. Calaminici, J.G. Gribben
Writing, review, and/or revision of the manuscript: R. Coutinho, A.J. Clear, A. Lee, M. Gomes da Silva, J. Cabeçadas, J.G. Gribben

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): A.J. Clear, J.G. Gribben, R. Coutinho
Study supervision: M. Calaminici, J.G. Gribben

Grant Support

R. Coutinho has a Doctoral grant from the Portuguese Foundation for Science and Technology (FCT) (SFRH/BD/68462/2010). This work was supported by program grants (to J.G. Gribben) P01 CA81538 from the National Cancer Institute and C1574/A6806 from Cancer Research UK.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 29, 2013; revised September 3, 2013; accepted October 2, 2013; published OnlineFirst October 11, 2013.

References

1. 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.
2. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proc Natl Acad Sci U S A 2003;100: 9991–6.
3. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, et al. Stromal gene signatures in large-B-cell lymphomas. N Eng J Med 2008;359: 2313–23.
4. Shaffer AL III, Young RM, Staudt LM. Pathogenesis of human B cell lymphomas. Annu Rev Immunol 2011;30:565–610.
5. Dunleavy K, Pittaluga S, Czuczman MS, Dave SS, Wright G, Grant N, et al. Differential efficacy of bortezomib plus chemotherapy within molecular subtypes of diffuse large B-cell lymphoma. Blood 2009;113: 6069–76.
6. Yang Y, Shaffer AL III, Emre NC, Ceribelli M, Zhang M, Wright G, et al. Exploiting synthetic lethality for the therapy of ABC diffuse large B cell lymphoma. Cancer Cell 2012;21:723–37.
7. Hernandez-Ilizaliturri FJ, Deeb G, Zinzani PL, Pileri SA, Malik F, Macon WR, et al. Higher response to lenalidomide in relapsed/refractory diffuse large B-cell lymphoma in nongerminal center B-cell-like than in germinal center B-cell-like phenotype. Cancer 2011;117:5058–66.
8. Wilson WH GJ, Gerecitano JF, Goy A, de Vos S, Kenkre VP, Barr PM. The Bruton's tyrosine kinase (BTK) inhibitor, ibrutinib (PCI-32765), has preferential activity in the ABC subtype of relapsed/refractory *de novo* diffuse large B-cell lymphoma (DLBCL): interim results of a multicenter, open-label, phase 2 study [abstract]. In: Proceedings of the 54th ASH Annual Meeting and Exposition; 2012 Dec 8–11. Atlanta, GA: Georgia World Congress Center. p. 120. Abstract nr 686.
9. 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.
10. Muris JJ, Meijer CJ, Vos W, van Krieken JH, Jiwa NM, Ossenkoppele GJ, et al. Immunohistochemical profiling based on Bcl-2, CD10 and MUM1 expression improves risk stratification in patients with primary nodal diffuse large B cell lymphoma. J Pathol 2006;208:714–23.
11. 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.
12. Nyman H, Jerkeman M, Karjalainen-Lindsberg ML, Banham AH, Leppa S. Prognostic impact of activated B-cell focused classification in diffuse large B-cell lymphoma patients treated with R-CHOP. Mod Pathol 2009;22:1094–1101.
13. 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.
14. 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.
15. Visco C, Li Y, Xu-Monette ZY, Miranda RN, Green TM, Tzankov A, 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.
16. Moskowitz CH, Zelenetz AD, Kewalramani T, Hamlin P, Lessac-Chenen S, Houldsworth J, et al. Cell of origin, germinal center versus nongerminal center, determined by immunohistochemistry on tissue microarray, does not correlate with outcome in patients with relapsed and refractory DLBCL. Blood 2005;106:3383–5.
17. 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.
18. Ott MM, Horn H, Kaufmann M, Ott G. The Hans classificator does not predict outcome in diffuse large B cell lymphoma in a large multicenter retrospective analysis of R-CHOP treated patients. Leuk Res 2012; 36:544–5.
19. Gutierrez-Garcia G, Cardesa-Salzmann T, Climent F, Gonzalez-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.
20. Fu K, Weisenburger DD, Choi WW, Perry KD, Smith LM, Shi X, et al. Addition of rituximab to standard chemotherapy improves the survival of both the germinal center B-cell-like and non-germinal center B-cell-like subtypes of diffuse large B-cell lymphoma. Clin Oncol 2008;26: 4587–94.
21. Berglund M, Thunberg U, Amini RM, Book M, Roos G, Erlanson M, et al. Evaluation of immunophenotype in diffuse large B-cell lymphoma and its impact on prognosis. Mod Pathol 2005;18:1113–20.
22. Cunningham D, Hawkes EA, Jack A, Qian W, Smith P, Mouncey P, et al. Rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisolone in patients with newly diagnosed diffuse large B-cell non-Hodgkin lymphoma: a phase 3 comparison of dose intensification with 14-day versus 21-day cycles. Lancet 2013;381:1817–26.
23. Nyman H, Adde M, Karjalainen-Lindsberg ML, Taskinen M, Berglund M, Amini RM, et al. Prognostic impact of immunohistochemically defined germinal center phenotype in diffuse large B-cell lymphoma patients treated with immunochemotherapy. Blood 2007;109:4930–5.
24. 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.
25. Seki R, Ohshima K, Fujisaki T, Uike N, Kawano F, Gondo H, et al. Prognostic impact of immunohistochemical biomarkers in diffuse large B-cell lymphoma in the rituximab era. Cancer Sci 2009;100:1842–7.
26. Thieblemont C, Briere J, Mounier N, Voelker HU, Cuccuini W, Hirchaud E, et al. The germinal center/activated B-cell subclassification has a prognostic impact for response to salvage therapy in relapsed/refractory diffuse large B-cell lymphoma: a bio-CORAL study. J Clin Oncol 2011;29:4079–87.
27. Montes-Moreno S, Roncador G, Maestre L, Martinez N, Sanchez-Verde L, Camacho FI, et al. Gcet1 (centerin), a highly restricted marker for a subset of germinal center-derived lymphomas. Blood 2008;111: 351–8.
28. Barrans SL, Fenton JA, Banham A, Owen RG, Jack AS. Strong expression of FOXP1 identifies a distinct subset of diffuse large B-cell lymphoma (DLBCL) patients with poor outcome. Blood 2004; 104:2933–5.

29. Natkunam Y, Zhao S, Mason DY, Chen J, Taidi B, Jones M, et al. The oncoprotein LMO2 is expressed in normal germinal-center B cells and in human B-cell lymphomas. Blood 2007;109:1636–42.
30. Barrans SL, Carter I, Owen RG, Davies FE, Patmore RD, Haynes AP, et al. Germinal center phenotype and bcl-2 expression combined with the International Prognostic Index improves patient risk stratification in diffuse large B-cell lymphoma. Blood 2002;99:1136–43.
31. van Imhoff GW, Boerma EJ, van der Holt B, Schuuring E, Verdonck LF, Kluin-Nelemans HC, et al. Prognostic impact of germinal center-associated proteins and chromosomal breakpoints in poor-risk diffuse large B-cell lymphoma. J Clin Oncol 2006;24:4135–42.
32. 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.
33. Mounier N, Briere J, Gisselbrecht C, Emile JF, Lederlin P, Sebban C, et al. Rituximab plus CHOP (R-CHOP) overcomes bcl-2–associated resistance to chemotherapy in elderly patients with diffuse large B-cell lymphoma (DLBCL). Blood 2003;101:4279–84.
34. Wilson KS, Sehn LH, Berry B, Chhanabhai M, Fitzgerald CA, Gill KK, et al. CHOP-R therapy overcomes the adverse prognostic influence of BCL-2 expression in diffuse large B-cell lymphoma. Leuk Lymphoma 2007;48:1102–9.
35. Salles G, de Jong D, Xie W, Rosenwald A, Chhanabhai M, Gaulard P, et al. Prognostic significance of immunohistochemical biomarkers in diffuse large B-cell lymphoma: a study from the Lunenburg Lymphoma Biomarker Consortium. Blood 2011;117:7070–8.
36. Iqbal J, Meyer PN, Smith LM, Johnson NA, Vose JM, Greiner TC, et al. BCL2 predicts survival in germinal center B-cell-like diffuse large B-cell lymphoma treated with CHOP-like therapy and rituximab. Clin Cancer Res 2011;17:7785–95.
37. Lawrie CH, Ballabio E, Soilleux E, Sington J, Hatton CS, Dirnhofer S, et al. Inter- and intra-observational variability in immunohistochemistry: a multicentre analysis of diffuse large B-cell lymphoma staining. Histopathology 2012;61:18–25.