Title: Poor concordance among nine immunohistochemistry classifiers of cell-of-origin for Diffuse Large B-cell Lymphoma: implications for therapeutic strategies

Rita Coutinho¹, Andrew James Clear¹, Andrew Owen¹², Andrew Wilson¹, Janet Matthews¹, Abigail Lee¹², Rute Alvarez³, Maria Gomes da Silva³, José Cabeçadas⁴, Maria Calaminici¹² and John G. Gribben¹

¹ Department of Hemato-Oncology, Barts Cancer Institute, Queen Mary University, London, United Kingdom
² Department of Histopathology, Barts Health NHS Trust, Royal London Hospital, London, United Kingdom
³ Department of Hematology and ⁴Department of Pathology, Portuguese Institute of Oncology Lisbon, Portugal

Running title: IHC cell-of-origin classifiers for DLBCL

Key words: Diffuse Large B-cell Lymphoma, cell-of-origin, immunohistochemistry, algorithms

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

Corresponding author:
John G. Gribben, Centre for Haemato-Oncology, Barts Cancer Institute - a Cancer Research UK Centre of Excellence, Queen Mary, University of London, John Vane Science Centre, Charterhouse Square, London EC1M 6BQ. Tel: +44 (0)20 7882 3805. Fax: +44 (0)20 7882 3891. Email: j.gribben@qmul.ac.uk

Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.
The authors have no conflict of interest to disclose

Text word count: 4374
Number of figures: 3 (1 in supplementary data)
Number of tables: 6 (2 in supplementary data)
IHC cell-of-origin classifiers for Diffuse Large B-cell Lymphoma

Statement of 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 (ABC) type 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 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.
Abstract

Purpose:
The opportunity to improve therapeutic choices on the basis of molecular features of the tumour 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 DLBCL. In order for targeted therapies to succeed in this disease, robust strategies that segregate patients into molecular groups with high reliability are needed. While 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 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% the tumors have been classified as GCB and 21% as ABC/non-GCB by all methods. None of the algorithms provided prognostic information in the R-CHOP 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.
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 biological segregation helps to explain the heterogeneous responses to standard treatment and provides a rational for investigation of novel targeted therapies. Emerging data support the notion that the two main DLBCL molecular groups, the activated B-like-(ABC) and the germinal centre B-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.

While 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 3 clinical trial (NCT01324596) aims at determining whether the addition of bortezomib to standard R-CHOP improves event-free survival and if 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, since its application is restricted to research purposes, there is presently a lack of standardized methodology for GEP analysis, which can lead to variable results both at the inter- and intra-laboratory level. This issue, which may impact on GEP results and on 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 co-workers (9) established the first IHC algorithm, with supposed high sensitivity for GEP classification. Subsequently, eight further strategies (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 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 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, a further 80 patients from the Portuguese Institute of Oncology, Lisbon 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 is detailed in Table 1 only for the R-CHOP treated cohort, for which outcome analysis was performed. No significant differences were observed regarding main initial features and outcome between patients with available tissue suitable for inclusion in the tissue microarrays (TMA) and the remainder (data not shown).

**Tissue Microarray (TMA) and Immunohistochemistry**

TMA were prepared from the paraffin blocks using a manual tissue arrayer (Beecher Scientific, Silver Spring, MD) in both institutions. Triplicate or duplicate 1-1.5 mm² 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 immunological 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 was 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 Hammamatsu Virtual Slide Scanner NanoZoomer 2.0 (Hammamatsu, Japan), 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 (RC, AL, MC) in order for scoring criteria to be standardized for each marker. Each case was scored as positive or negative according to the cutpoints defined in the original publications and detailed in Figure 1. It is important to stress that this means that some antibodies were analyzed for more than one cutpoint, 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 non-concordant results, the core with highest large cell infiltration was used. After each antibody has been 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 experienced hematopathologist (MC). 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 interpatient differences in staining intensity.

Statistical analysis

In order to render outcome analysis more relevant for the clinical community, all results regarding 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 chi-square or Fisher's exact test, when appropriate. A variable number of cases failed to yield reliable staining results for the reasons already detailed and were excluded from 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 Kappa statistics.

In the univariate analysis, log rank tests were performed. All parameters of the 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-value ≥ 0.2 on univariate analysis.

The outcomes, measured from date of diagnosis to occurrence of event or date of last follow-up, were: overall survival (OS), the event being death from any cause and event free survival (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, Chicago, IL) and Prism version 5.03 (GraphPad Software, La Jolla, CA).

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 appears 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 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’s 0.311, p<0.001) and GCET1 (Pearson’s 0.45, p<0.001), and a negative correlation with the post-GC marker MUM1 (Pearson’s -0.164, p=0.05).

Both GCB and ABC DLBCLs can harbor genetic aberrations involving BCL6, leading to protein over-expression. Nevertheless, BCL6 is considered a GC marker. BCL6 expression was comparable to previous reports (9,19,25,26). A positive correlation was observed with other GCB-related proteins (CD10; GCET1, Pearson’s 0.39, p<0.001; and LMO2, Pearson’s 0.51, p<0.001) and the ABC marker FOXP1 (Pearson’s 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 Choi’s data (13). Moreover, this higher cutoff improved correlation with other post-GC markers, such as FOXP1 (Pearson’s 0.19 to 0.37, p<0.01). A significant correlation between BCL2 and MUM1 expression was also documented (Pearson’s 0.36 to 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 cutpoint of 80%, and this only increased to 20% at the lower cutoff of 30%. At both cutpoints, 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’s 0.34, p<0.001), and negative correlation with MUM1 (Pearson’s -0.19, p=0.03) were detected, with weaker correlations at the higher cutpoint.
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 >30% and 72% had >60% of FOXP1+ cells. At a higher cut-off 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 GC B-cells (29) in lymph nodes and to a subset of GCB-DLBCLs. Natkunam et al (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 appears 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 utilized 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. Hans's 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 GCB patients from the Hans method and a further eight patients scored CD10+/MUM1-.
The Choi classifier (13) relies on the expression of three antigens (GCET1, MUM1 and CD10) for initial allocation and on a further two (BCL6 and FOXP1) for final decision. Cutoff points were adopted according to appropriate detection of a single series molecularly profiled patients (13). Using the cutoffs defined in their paper, 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 et al (14) proposed the modified Choi (Fig. 1), which allocated 35 (25%) and 105 of our patients to GCB and non-GCB groups, respectively. Compared to the Choi algorithm, the modified version reclassified both GCB (64 cases) and non-GCB cases (8 cases). Due to 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 which is highly sensitive at identifying ABC patients, the subgroup that may be more amenable to targeting with new agents. Nyman et al (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 ABC cases (110, 78%) were classified as ABC. In 34 cases LMO2 expression was used for decision, being positive in half of the cases.
Finally Visco et al (15) recently launched another method with an increased overlap with GEP data. Similarly to Hans, CD10 plays the central role for GCB allocation (41 of the 53 GCB patients were CD10"). Regarding 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 pair wise agreement using the general kappa statistics, a method which tests for interscoring reliability. The kappa is considered a robust measure since it takes into account the agreement occurring by chance. Fig. 2 illustrates the strength of agreement among all scoring systems. Poor and fair kappa values were detected in 44.4% on pair wise concordance assessment; and in only 20% was kappa excellent or 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 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 value in predicting outcome in the R-CHOP cohort. 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% cutpoint) 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 to 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 over-simplified 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 underway 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 utilizing GEP for this purpose will clarify the utility of targeted therapies in the clinical trial 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 classification of DLBCL subsets, would be readily applicable in clinical practice and would be incorporated into diagnostic workup within Hematopathology clinical laboratories. However, based upon previous work and the data presented here, we suggest that much work needs to be done to standardize the Hans and other IHC methods which 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. Whereas this study would have been enhanced by the availability of GEP as a “gold-standard”, the methodology employed here does not require such a comparison, since 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 others (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 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 end-point. Moreover, results were made more robust by using the kappa statistics, which takes into account the agreement occurring by chance. Scoring allocation appeared 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 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 B-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, while 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 >30% expression of MUM1 in a higher proportion of patients than previous studies and this cannot be explained by commonly reported reasons such as non-specific cytoplasmic background staining and target cell artifacts (24). Choi et al (13) claimed that a higher cutoff level of expression of 80% was required to achieve high specificity for ABC cases. Using this cutpoint, 42% of our patients were deemed positive for MUM1 expression, in keeping with Choi’s data. This, however, highlights the difficulties of standardizing results based upon 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 Choi and Visco classifiers, the ABC subset was enriched for older patients. This suggests that clinical prognostic factors might interact with biological 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 upon 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, since 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 render 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.

References


### Table 1: Clinical features of the R-CHOP series

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>median (range)</td>
<td>62 (16-86)</td>
</tr>
<tr>
<td>&gt;60 years</td>
<td>81 (55)</td>
</tr>
<tr>
<td>Male sex</td>
<td>77 (52)</td>
</tr>
<tr>
<td>B-symptoms</td>
<td>38 (26)</td>
</tr>
<tr>
<td>≥2 extranodal sites</td>
<td>24 (16)</td>
</tr>
<tr>
<td>Ann Arbor stage III-IV</td>
<td>81 (55)</td>
</tr>
<tr>
<td>ECOG performance status ≥ 2</td>
<td>23 (16)</td>
</tr>
<tr>
<td>High LDH</td>
<td>97 (66)a</td>
</tr>
<tr>
<td>IPI (Low / Low-int / High-int/High)</td>
<td>54 (37)/ 33 (23)</td>
</tr>
<tr>
<td>Treatment response</td>
<td></td>
</tr>
<tr>
<td>CR</td>
<td>110 (75)</td>
</tr>
<tr>
<td>PR</td>
<td>22 (15)</td>
</tr>
<tr>
<td>SD/PD</td>
<td>15 (10)</td>
</tr>
<tr>
<td>Mortality</td>
<td>47 (32)</td>
</tr>
<tr>
<td>Follow up (months, range)</td>
<td>53 (4 - 71)</td>
</tr>
</tbody>
</table>

Clinical characteristics were available for 147 patients. Significant differences between the two R-CHOP groups are detailed in the results. a a significantly higher number of patients diagnosed at IPO had high LDH at diagnosis by Fishers exact test. n: number; ECOG: Eastern Cooperative Oncology Group; LDH: lactate dehydrogenase; IPI: International Prognostic Index; Low-int: low-intermediate; High-int: high-intermediate; CR: complete response; PR: partial response; SD: stable disease; PD: progressive disease; NS: not significant.
### Table 2: Single marker analysis

<table>
<thead>
<tr>
<th>Antigens</th>
<th>positive [n (%)]</th>
<th>negative [n (%)]</th>
<th>unclassifiable (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD10 (30%)</td>
<td>41 (28)</td>
<td>106 (72)</td>
<td>4</td>
</tr>
<tr>
<td>BCL6 (30%)</td>
<td>86 (61)</td>
<td>54 (39)</td>
<td>11</td>
</tr>
<tr>
<td>MUM1 (30%)</td>
<td>109 (76)</td>
<td>35 (24)</td>
<td>7</td>
</tr>
<tr>
<td>MUM1 (80%)</td>
<td>60 (42)</td>
<td>84 (58)</td>
<td>7</td>
</tr>
<tr>
<td>GCET1 (30%)$^a$</td>
<td>29 (20)</td>
<td>118 (80)</td>
<td>4</td>
</tr>
<tr>
<td>GCET1 (80%)</td>
<td>12 (8)</td>
<td>135 (92)</td>
<td>4</td>
</tr>
<tr>
<td>FOXP1 (30%)</td>
<td>116 (82)</td>
<td>25 (18)</td>
<td>10</td>
</tr>
<tr>
<td>FOXP1 (60%)</td>
<td>102 (72)</td>
<td>39 (28)</td>
<td>10</td>
</tr>
<tr>
<td>FOXP1 (80%)$^b$</td>
<td>77 (55)</td>
<td>64 (45)</td>
<td>10</td>
</tr>
<tr>
<td>LMO2 (30%)</td>
<td>79 (56)</td>
<td>62 (44)</td>
<td>10</td>
</tr>
<tr>
<td>BCL2 (50%)</td>
<td>96 (69)</td>
<td>43 (31)</td>
<td>12</td>
</tr>
</tbody>
</table>

Absolute number and percentage of classifiable cases is highlighted for each antigen. Staining and analysis were performed centrally. $^a$ a significantly higher proportion of GCET1$^+$ cases (30% cut-off) 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 the cutoff of 80% (p=0.001). n: number; NS: not significant.

### Table 3: Distribution of R-CHOP treated patients according to the nine IHC classifiers

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

Unclassifiable cases were excluded for percentage calculation. Significant differences observed in the cohorts from the two Institutions according to Fisher’s exact test are detailed in the text.
Table 4: Survival analysis according to clinical characteristics, IPI and cell-of-origin IHC classifiers

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

OS: overall survival; EFS: event-free survival; ECOG: Eastern Cooperative Oncology Group; LDH: lactate dehydrogenase; IPI: International Prognostic Index; Low-int: low-intermediate; High-int: high-intermediate; NS: not significant

FIGURE LEGENDS

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; (I) Visco-Young.

Figure 2: Pair wise agreement according to kappa statistics.
*modified
Figure 1

A

CD10 > 30%

+ GCB

- BCL6 > 30%

+ non-GCB

- MUM1 > 30%

- GCB

- non-GCB

B

CD10 > 30%

+ GCB

- MUM1 > 30%

- ABC

- FOXP1 > 80%

- OTHER

- ABC

C

MUM1 > 30%

+ GCB

- non-GCB

D

BCL2 > 50%

+ GCB

- CD10 > 30%

- non-GCB

E

+ GCET1 > 80%

- MUM1 > 80%

+ CD10 > 30%

+ non-GCB

- GCB

- GCB

- BCL6 > 30%

- ABC

- FOXP1 > 80%

- GCB

F

FOXP1 > 80%

+ GCET1 > 80%

- CD10 > 30%

+ GCB

- ABC

- ABC

- GCB

- MUM1 > 30%

G

GCB

ABC

CD10 > 30%

+ VS = score 0-2

GCB

ABC

MUM1 > 30%

+ VS = score 0-2

GCB

ABC

FOXP1 > 80%

H

LMO2 > 30%

+ GCB

- ABC

- GCB

- ABC

I

CD10 > 30%

+ GCB

- ABC

- non-GCB

- BCL6 > 30%

- non-GCB
Figure 2

<table>
<thead>
<tr>
<th>Kappa</th>
<th>Hans</th>
<th>Hans*</th>
<th>Nyman</th>
<th>Choi</th>
<th>Choi*</th>
<th>Natkunam</th>
<th>Tally</th>
<th>Muris</th>
<th>Visco</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hans*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nyman</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choi*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natkunam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tally</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muris</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visco</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Poor</th>
<th>Fair</th>
<th>Moderate</th>
<th>Excellent</th>
<th>Very good</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Poor concordance among nine immunohistochemistry classifiers of cell-of-origin for Diffuse Large B-cell Lymphoma: implications for therapeutic strategies

Rita Coutinho, Andrew Clear, Andrew Owen, et al.

Clin Cancer Res  Published OnlineFirst October 11, 2013.

Updated version  Access the most recent version of this article at: doi:10.1158/1078-0432.CCR-13-1482

Supplementary Material  Access the most recent supplemental material at: http://clincancerres.aacrjournals.org/content/suppl/2013/10/09/1078-0432.CCR-13-1482.DC1

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.