VOLUME 25 · NUMBER 7 · MARCH 1 2007

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

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

Daphne de Jong, Andreas Rosenwald, Mukesh Chhanabhai, Philippe Gaulard, Wolfram Klapper, Abigail Lee, Birgitta Sander, Christoph Thorns, Elias Campo, Thierry Molina, Andrew Norton, Anton Hagenbeek, Sandra Horning, Andrew Lister, John Raemaekers, Randy D. Gascoyne, Gilles Salles, and Edie Weller

From the Netherlands Cancer Institute; Academic Medical Center, Amsterdam; University Medical Center Nijmegen, the Netherlands; Institute of Pathology, University of Würzburg, Würzburg; Department of Pathology, Hematopathology Section, University Hospital Schleswig-Holstein, Campus Kiel; University Clinic Schleswig-Holstein, Campus Luebeck, Germany; Department of Pathology & Medical Oncology, British Columbia Cancer Agency, University of British Columbia, Vancouver, Canada; Department of Pathology, Inserm U617, Hôpital Henri Mondor, Créteil; Université Paris-Descartes; AP-HP, Hôtel-Dieu, Paris; Hospices Civils de Lyon & Université Claude Bernard Lyon-1, Lyon, France; CR-UK Medical Oncology Unit, St Bartholomew's Hospital, London, United Kingdom; Karolinska Institutet, Stockholm, Sweden; Hospital Clinic, University of Barcelona, Barcelona, Spain; Stanford University Medical Center, Palo Alto, CA; and Dana-Farber Cancer Institute, Boston, MA.

Submitted October 5, 2006; accepted December 7, 2006.

Supported by the van Vlissingen Lymphoma Foundation. In addition, unrestricted grants were received from Genentech, Millennium Pharmaceuticals Inc, Roche International, and Schering AG.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Address reprint requests to Daphne de Jong, MD, PhD, Department of Pathology, the Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, the Netherlands; e-mail: d.d.jong@nki.nl.

0732-183X/07/2507-805/$20.00

DOI: 10.1200/JCO.2006.09.4490

ABSTRACT

Purpose
The results of immunohistochemical class prediction and prognostic stratification of diffuse large B-cell lymphoma (DLBCL) have been remarkably various thus far. Apart from biologic variations, this may be caused by differences in laboratory techniques, scoring definitions, and inter- and intraobserver variations. In this study, an international collaboration of clinical lymphoma research groups from Europe, United States, and Canada concentrated on validation and standardization of immunohistochemistry of the currently potentially interesting prognostic markers in DLBCL.

Patients and Methods
Sections of a tissue microarray from 36 patients with DLBCL were stained in eight laboratories with antibodies to CD20, CD5, bcl-2, bcl-6, CD10, HLA-DR, MUM1, and MIB-1 according to local methods. The study was performed in two rounds firstly focused on the evaluation of laboratory staining variation and secondly on the scoring variation.

Results
Different laboratory staining techniques resulted in unexpectedly highly variable results and very poor reproducibility in scoring for almost all markers. No single laboratory stood out as uniformly poor or excellent. With elimination of variation due to staining, high agreement was found for CD20, HLA-DR, and CD10. Poor agreement was found for bcl-6 and Ki-67. Optimization of techniques and uniformly agreed on scoring criteria improved reproducibility.

Conclusion
This study shows that semiquantitative immunohistochemistry for subclassification of DLBCL is feasible and reproducible, but exhibits varying rates of concordance for different markers. These findings may explain the wide variation of biomarker prognostic impact reported in the literature. Harmonization of techniques and centralized consensus review appears mandatory when using immunohistochemical biomarkers for treatment stratification.

J Clin Oncol 25:805-812. © 2007 by American Society of Clinical Oncology

INTRODUCTION

The use of immunohistochemical methods has become part of the routine diagnostic procedure in several malignancies, and is essential in lymphoma. In the last 10 years, markers have been identified that influence a patient's prognosis. This has led to the proposed use of these markers for risk stratification of lymphoma patients and to the development of specific therapeutic strategies. Since the recognition of two biologic subtypes of diffuse large B-cell lymphoma (DLBCL) on the basis of gene-expression profiling,[1,2] the exploration of the clinical relevance of this subtyping has been the subject of many studies. The prognostic stratification of the germinal center B-cell–like (GCB) and activated B-cell–like (ABC) has been reproducible in most gene-expression studies by different groups.[3,4] To enable implementation of prognostic stratification as a basis for treatment choice in clinical practice, however, more broadly applicable methods are needed.

Immunohistochemistry using a limited number of markers is an attractive alternative technique that enables exploration in larger retrospective and prospective studies, in uniformly treated series from

clinical trials, and in more rare and specific patient populations. Moreover, integration of many other markers that have been identified over the years as single prognostic markers in DLBCL is feasible. Several groups of authors have used similar approaches to translate the biologic information of the GCB- versus non–GCB-like subtypes in sizable patient series.[5-11] Although the markers and algorithms were highly similar, the results were remarkably various. Some of these groups found a significant prognostic value using the immunohistochemical class prediction similar to the gene-expression method, whereas others did not. The inconsistency in results not only applies to the rather complex biologic subtype stratification, but similarly holds true for single immunohistochemical prognostic markers such as bcl-2, bcl-6, survivin, FoxP1, and Ki-67 in DLBCL.[12-21]

Several biologic and technical causes explaining the inconsistent findings for the prognostic value of markers in DLBCL may be proposed: selection of specific patient series (specific age groups, relative contribution of nodal versus extranodal disease); treatment factors (nonuniform treatment ± rituximab); laboratory technical variations (such as various antigen retrieval and signal amplification techniques); scoring criteria and definitions; and inter- and intraobserver variations.

In October 2003, the Lunenburg Lymphoma Biomarker Consortium (LLBC) was instituted as an international collaboration of nine leading clinical lymphoma research groups from Europe, United States, and Canada to unite their efforts in translational research.[22] The main research topic was to determine those biomarkers believed to be important for prognosis in DLBCL. The aims of this project are to standardize the measurement of biomarkers in DLBCL and to validate the prognostic relevance of important markers in large clinical trials performed by cooperative groups throughout the world. Before launching a comprehensive study on biopsy samples from patients treated in clinical trials, the group concentrated on validation and standardization of immunohistochemistry of the currently interesting prognostic markers in DLBCL. This effort focused on the evaluation of technical and interobserver variation in the context of strict scoring criteria and definitions.

PATIENTS AND METHODS

Tissue Microarray Construction

Tissue microarrays (TMAs) were prepared at the Department of Pathology of the British Columbia Cancer Center (Vancouver, Canada) from 36 representative patient samples with DLBCL and two tonsil samples with adequate archival formalin-fixed and paraffin-embedded material retrieved from six different laboratories and collected between 1984 and 2004. Representative 0.6-mm cores were taken and re-embedded in duplicate per recipient block. Six identical recipient blocks were constructed. Five-microliter sections were then cut from each TMA in eight laboratories and stained with antibodies to CD20, CD5, bcl-2, bcl-6, CD10, HLA-DR, MUM1, and MIB-1 according to local methods (Table 1).

Criteria and Scoring Methods for Immunohistochemistry

Each core was evaluated for percentage of tumor cells stained by visual estimation and the maximum of the two cores was recorded. The LLBC pathologists convened twice to determine and refine the criteria for scoring before the first rotation round. The scoring categories and requirements for internal controls are listed in Table 2. The Ki67 staining from laboratory 5 and the MUM1 staining from laboratory 4 were not considered due to suboptimal technical results precluding evaluation. Laboratory 3 did not perform a HLA-DR staining. In the first rotation round, the set of eight immunohistochemical stains from each laboratory was scored by the local pathologist and two other pathologists to assess staining and scoring variation. Therefore, for each patient, 24 scores are generated.

Based on the results from the first rotation round, which showed significant staining effects, the LLBC pathologists convened a meeting to discuss the results and to compare directly all available stained TMA slides. The optimal staining per marker was selected in terms of expected best scoring reproducibility, occurrence of minimal artifacts, and the representation of the expected biologic range/variation of the marker in DLBCL. Scoring variation was evaluated further in a second rotation round in which all nine pathologists scored the optimal set of stainings. For CD5 and bcl-6, two stains were selected that differed in staining characteristics, precluding interpretation on expected best scoring reproducibility. For CD5, identical scoring criteria were applied on both stains. For bcl-6, two different sets of scoring criteria were used: one based on cell percentages and one based on staining intensity. A scoring manual was constructed as an additional guideline.

Table 1. Primary Antibodies and Protocols

Antibody and Method	Laboratory No. 1	2	3	4	5	6	7	8
Ki67	DAKO* MIB1	DAKO MIB1	DAKO MIB1	DAKO MIB1	Locally produced	DAKO MIB1	DAKO MIB1	DAKO MIB1
CD20	DAKO L26	DAKO L26	DAKO L26	DAKO L26	Locally produced	DAKO L26	DAKO L26	DAKO L26
CD5	Novacastra† 4C7	Novacastra 4C7	Novacastra 4C7	Novacastra 4C7	MEDAC‡ NCL-CD5 4C7	Novacastra 4C7	Novacastra 4C7	Novacastra 4C7
bcl-2	DAKO 124	DAKO 124	DAKO 124	DAKO 124	ZYMED§ bcl-2-100	BIOCARTA‖ 100D5	DAKO 124	DAKO 124
CD10	Novacastra 56C6	Novacastra 56C6	Novacastra 56C6	Novacastra 56C6	MEDAC CD10 270	Novacastra 56C6	Novacastra 56C6	Novacastra 56C6
bcl6	DAKO PG-B6p	DAKO PG-B6p	DAKO PG-B6p	DAKO PG-B6p	DAKO PG-B6p	DAKO PG-B6p	DAKO PG-B6p	DAKO PG-B6p
MUM1	DAKO MUM1p	DAKO MUM1p	DAKO MUM1p	DAKO MUM1p	DAKO MUM1p	DAKO MUM1p	DAKO MUM1p	DAKO MUM1p
HLA-DR	DAKO TAL.1B5	DAKO TAL.1B5	Not done	DAKO TAL.1B5	DAKO TAL.1B5	DAKO TAL.1B5	DAKO TAL.1B5	DAKO TAL.1B5
Method	Automatic	Automatic	Automatic	Automatic	Manual	Automatic	Automatic	Manual
Development	ABC/DAB	Envision*/DAB	Powervision¶	Biotin/streptavidin	APAAP	ABC/DAB	ChemMate*/DAB	ABC/DAB

Abbreviations: ABC, activated B-cell–like; DAB, diaminobenzidine; APAAP, alkaline phosphatase anti-alkaline phosphatase.
*DAKO, Glostrup, Denmark.
†Novacastra, Newcastle upon Tyne, United Kingdom.
‡MEDEC, Wedel, Germany.
§ZYMED, San Francisco, CA.
‖Biocarta, San Diego, CA.
¶Immunovision Technologies, Duiven, the Netherlands.

Table 2. Scoring Criteria for Immunohistochemistry in the Second Rotation Round

Antibody	Scoring Criteria
CD20	Score as positive/negative in tumor cells
CD5	No staining, 1%-25%, 26%-50%, 51%-75%, > 75%; for the score designated no staining, an internal staining control must be present; T cells serve as internal controls
HLA-DR	Score as positive/negative in tumor cells; for the score designated as negative, an internal staining control must be present; reactive small B cells and T cells and accessory cells serve as internal controls
CD10	Score as positive/negative in tumor cells; for the score designated as negative, an internal control must be present; granulocytes and stromal fibroblasts serve as internal controls
bcl-2	No staining (0%-5%), 5%-25%, 26%-50%, 51%-75%, > 75%; for the score designated as no staining, an internal staining control must be present; staining intensity is scored as weak (weaker than internal T cells) and strong (equal or stronger than internal T cells)
Ki67	No staining, 1%-25%, 26%-50%, 51%-75%, 76%-95%, > 95%; for the score designated as no staining, an internal staining control must be present
MUM1	No staining (0%-5%), 5%-25%, 26%-50%, 51%-75%, > 75%; for the score designated as no staining, an internal staining control must be present; activated T cells serve as internal controls
bcl-6, laboratory 3	No staining (0%-5%), 5%-25%, 26%-50%, 51%-75%, > 75%; for the score designated as no staining, an internal staining control must be present; internal controls may be sparse and consist of T cells
bcl-6, laboratory 7	Strong (saturated), strong variable (variable with strong to moderate staining variation), variable weak (variation between weak and moderate staining intensities), and weak (negative with sporadically staining cells); for the score designated as no staining, an internal staining control must be present; internal controls may be very sparse and consist of T cells

Statistical Analysis

Metrics used to evaluate agreement included overall agreement between pairs of laboratories as well as the proportion of patients for whom all scores agree. The overall pair-wise agreement was adjusted for the expected proportion of agreement assuming the scoring laboratories were independent using the generalized κ statistics.[23] The level of agreement for the κ statistic was evaluated based on the following ranges: less than 0, poor; 0.0 to 0.2, slight; 0.2 to 0.4, fair; 0.4 to 0.6, moderate; 0.6 to 0.8, substantial; and 0.8 to 1.0, almost perfect. The SE of the generalized κ statistic was estimated using the bootstrap method with 2,000 replications.[24] Resampling was performed at the patient level to conserve the correlation structure of the scores within a patient. The bootstrap CIs were computed based on the percentiles of the bootstrap distribution of the statistic.[24] The agreement metrics were evaluated including or excluding the category that was not scored.

We also evaluated whether combining biologically homogeneous categories could improve agreement. This analysis was performed for bcl-6, CD5, and Ki-67 in the second rotation round. Using the most generally applied algorithm to distinguish GCB versus non-GCB DLBCL on the basis of immunohistochemistry for CD10, bcl-6, and MUM-1, agreement and generalized κ statistics among laboratories was performed for data from both rotation rounds. A cutoff level of 25% was used for each marker as a positive score in the GCB versus non-GCB classification.

RESULTS

Agreement Across Staining and Scoring Laboratories

Agreement results are summarized for the first and second rotation round in Tables 3 and 4, respectively, and in Figure 1. The majority of the markers show a large difference in agreement between both rounds, with improvement in the second round in the overall pairwise agreement of 18% or higher. Only CD20 stands out as a uniformly reliably marker to score in all situations. HLA-DR was reproducible in terms of staining or scoring, with agreement ranging between 86% and 92% for all staining laboratories except one in the first round, and agreement of 95% in the second round.

CD10 may be considered as a reliably scored marker. However, reproducibility comes at the cost of a high percentage of nonassessable patients due to the absence of a positive internal control for suboptimal stains, as reflected by the poor agreement results with the inclusion of the patients who were not scored in the first round (pairwise agreement of 65% *v* 87% with and without the patients who were not scored, respectively) with improvement in the second round (pairwise

Table 3. Agreement Percentages and the Generalized κ Statistic From the First Rotation Round Combined Across Staining Laboratories

Marker	Percent Agreement of 24 scores*	Agreement in Pairs of Laboratories		κ Statistic	
		%	Bootstrap SE	Generalized	Bootstrap SE
BCL2	0	47	7	0.23	0.07
BCL6	0	34	3	0.17	0.03
CD10	3	65	4	0.39	0.08
CD20†	79	95	3	—	—
CD5	18	79	4	0.25	0.16
HLA-DR‡	0	77§	4	0.32	0.09
Ki67‡	0	35	3	0.14	0.04
MUM1‡	0	34	4	0.16	0.05
GCB	3	57	5	0.36	0.07

*Computed as the percent of the 36 patients among whom the 24 scores (three pathologists scored eight stains per patient) agree.
†κ statistic not computed for CD20 because 99.6% of the patients scored are positive, resulting in high expected agreement and small denominator of the κ statistic.
‡One staining laboratory for Ki67, one staining laboratory for MUM1 and GCB, and one staining laboratory for HLA-DR were excluded.
§Pairwise agreement was high for seven of the eight staining laboratories (86%-92%) and low for one staining laboratory (34%).

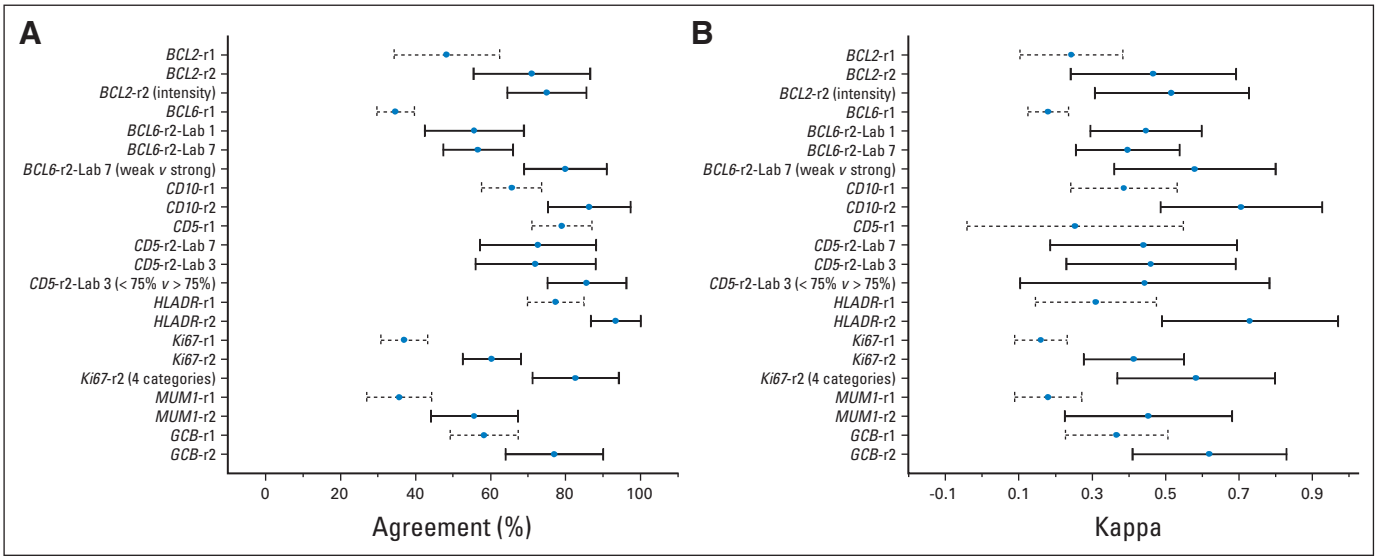

Fig 1. Percent agreement (A) and generalized κ statistic (B) and the 95% bootstrap percentile CIs from the first round (denoted by r1 in the labels and [– – –] in the figure) and the second round (denoted by r2 in the labels and [——] in the figure).

agreement of 87% *v* 95% with and without the patients who were not scored, respectively).

CD5 and bcl-2 form a class of markers that pose more problems. CD5 staining was strongly influenced by technical variations. Using clone 4C7 with standard ABC/diaminobenzidine gives distinctly different results (pair-wise agreement between 91% and 98%) than visualization with maximized enhancement systems (pairwise agreement between 68% and 69%; ChemMate Detection Kit; DAKO and Powervision, Immunovision Technologies, Duiven, the Netherlands) with far stronger membranous staining at the cost of increased intracytoplasmic background staining. Experience with flow results indicates that this staining is actually specific and should be considered as positive (B. Sander and R.D. Gascoyne, personal communication, November 2005). Therefore, for the second rotation round both maximized stains were included. However, the second round results show that extreme enhancement introduces an unacceptable level of background staining, which caused a high percentage of patients who could not be scored (11% not scored; Figs 2A and 2B). CD5 was initially considered in five scoring categories. The distribution of patients over these categories, with 9% of the scores in the intermediate categories, suggested that the biologic dichotomy could be placed at a single higher level ($\geq$ 75%; Figs 2A and 2B). This approach improved the interobserver agreement, but agreement remained at a moderate level per the κ statistic.

Bcl-2 showed only fair agreement in the first rotation round due to staining variations (47% pairwise agreement; $\kappa = 0.23$) and a relatively high percentage of patients who could not be scored in some of the stains (3% to 35%). In the second round with the optimal stain, only 1% of samples could not be scored and moderate agreement could be reached (70% pairwise agreement; $\kappa = 0.45$). As an alternative approach, the intensity of cytoplasmic staining in tumor cells compared with reactive T cells in the same sample was considered. This could be performed with somewhat better reproducibility (74% pairwise agreement; $\kappa = 0.51$). Whether this feature is of biologic and prognostic relevance remains to be studied in the context of a clinical series, however.

Despite the use of freshly cut sections, the nuclear markers BCL-6, MUM-1, and Ki-67 proved to be the markers most influenced by extreme laboratory variations, resulting in generalized κ scores below 0.2. For MUM-1, scoring variation was improved in the second rotation round. When all pathologists scored the same stain, a moderate agreement (54% pairwise agreement; $\kappa = 0.41$) could be reached.

When eliminating the laboratory staining variation, the reproducibility of the scoring for Ki-67 improved in the second round, but the separation of the two higher categories (76% to 95% and $>$ 95%) was highly variable (percent of patients $>$ 95% ranged from 15% to 52% for the nine scoring laboratories; Fig 2C). Indeed, when considering these two categories together in a four categories score, a moderate agreement ($\kappa = 0.58$) could be reached and the percent of patients with scores more than 75% was less variable (66% for one laboratory and between 73% and 79% for eight scoring laboratories).

Bcl-6 was found to be the most variable and most difficult marker to score. Despite use of the same primary antibody, the staining results varied dramatically. Different laboratory techniques were found to influence strongly the level of sensitivity of the staining (Figs 2D and 2E). Two laboratories produced positive staining with bcl-6 in virtually all patients with DLBCL, which is in line with expression data of RNA-based techniques. Generally, the staining is weaker, reflecting a less sensitive technique. Two different scoring systems were used for these two patterns; based on intensity and based on relative percentages of positive cells. The classical method based on percentages resulted in moderate agreement (53% pairwise agreement; $\kappa = 0.42$), but at the cost of an unacceptably high percentage of patients who could not be scored (15%). Intensity scoring also yielded moderate agreement (80% pairwise agreement; $\kappa = 0.58$) when dichotomized in simplified weak versus strong categories.

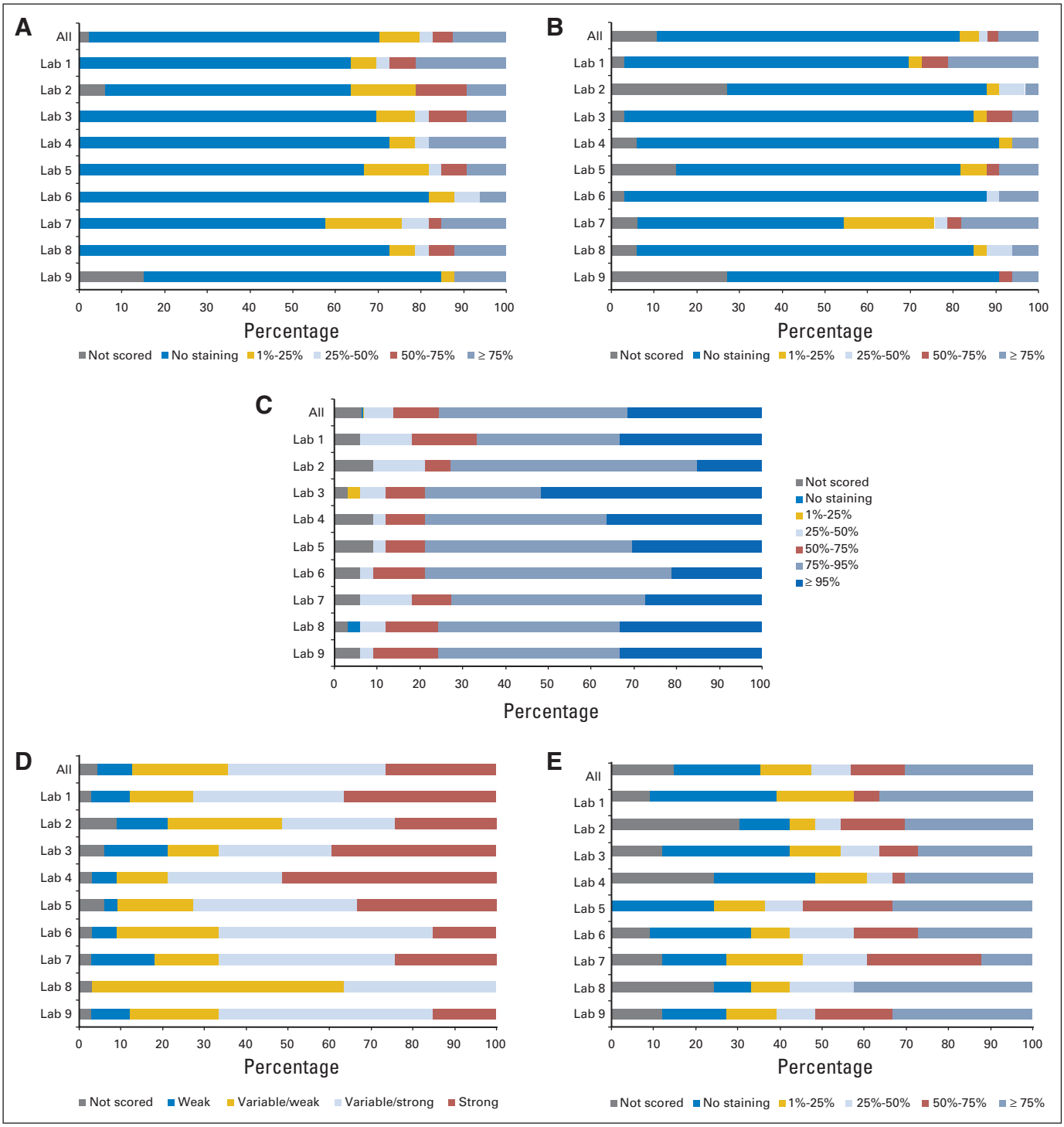

Fig 2. Distribution of the scores of the nine pathologists from the second rotation round for (A and B) CD5, (C) Ki-67, and (D and E) bcl-6. The *x* axis is the percentage of the scores in each category.

Reproducibility of GCB- Versus ABC-Like DLBCL Class Assignment

The combined analysis of CD10, bcl-6, and MUM-1 according to set algorithms may be a surrogate for the gene expression signatures of the prognostically relevant classes of ABC- versus GCB-like DLBCL. Staining and scoring variations may have a direct effect on the reproducibility of the immunologic class assignment. From the second rotation round with the optimal stains, pairwise agreement of 77% with a κ value of 0.62 could be reached. However, exclusion of patients who could not be scored for one or more of the relevant markers according to the set criteria (mostly lack of internal control), resulted in

pairwise agreement of 89% with a κ value of 0.77. As expected, staining variations resulted in lower agreement (Table 4).

DISCUSSION

The modern approach to cancer treatment is more and more driven by biologic insights aimed at tailored therapy. Therefore, more demand than ever is put on the pathologist to provide reproducible and reliable information on markers and biologic subclassifications. Before launching a comprehensive study on biopsy samples from DLBCL patients treated in clinical studies, this LLBC study explored variations introduced by laboratory techniques, interobserver variations, and scoring reproducibility of a set of established and potentially important immunohistochemical markers for DLBCL. Even though all stainings were performed in experienced laboratories with a special focus on hematopathology, laboratory variations had a major impact on levels of agreement. When the staining variation was eliminated, scoring proved to be highly reproducible between pathologists for several markers such as CD20, CD10, and HLA-DR, and sufficiently so for MUM-1, bcl-2, and CD5. However, for other markers, including Ki-67 and bcl-6, the reproducibility was at a lower level even with exclusion of the staining variation. Importantly, however, the results show that when optimal-quality staining is used, the overall reproducibility and the agreement on classification as ABC- versus GCB-like DLBCL can be improved significantly, especially when strict scoring guidelines are followed.

These results provide important insights for the variable conclusions in the currently available literature on prognostic markers in DLBCL. Bcl-2 generally has been considered as a consistent prognostic marker in many series of DLBCL, with an independent prognostic value for shorter disease-free survival and overall survival in patients with bcl-2–positive DLBCL. However, a critical look at the various results during the last 10 years shows considerable variation.[5-7,9,12-15,25] Although a predictive value for DFS is reported in most series, the predictive value of bcl-2 positivity for overall survival is not reliably consistent among those reports. Although some series may lack the statistical power to detect survival difference, variations in scoring and interpreting bcl-2 staining are likely to account for such discrepancies. Indeed, the percentage of DLBCL considered as positive in these series varies between 24% and 88%, with cutoff levels varying between 10% and 60%. Moreover, this validation study shows that bcl-2 staining is strongly influenced by local technical aspects. As long as these aspects are unresolved, it is difficult to define the relevant factors that determine the prognostic value of bcl-2 in biologic subgroups[26,27] or in rituximab-treated patients[25] outside a centralized and validated immunohistochemical approach and pathologic review.

This study shows that agreement was better for markers scored with only two categories compared with multiple categories. This would form a strong argument against too refined scoring and for omitting essentially nonreproducible cutoff points in situations that would be permitted from a biologic point of view. For daily practice, this may have implications for the distinction of DLBCL, Burkitt-like, and atypical Burkitt lymphoma, in which a high proliferation rate is one of the defining parameters. We could not reproducibly score Ki-67, however, in categories of 75% to 95% versus more than 95% ($\kappa = 0.39$), whereas combining the highest categories improved reproducibility dramatically ($\kappa = 0.58$), showing that the problem indeed lies in the upper ranges and that 95% of positive cells may not be a reliable cutoff point. This information indicates that Ki-67 may not be a marker of choice for the classification of Burkitt(-like) lymphoma in daily practice.

Modern immunohistochemical enhancement techniques (standard citrate retrieval, Powervision; Immunovision Technologies, Duiven, the Netherlands; Envision; DAKO, Glostrup, Denmark) have greatly increased the detection levels of proteins. In general, this can be

Table 4. Agreement Percentages and the Generalized κ Statistic From the Second Rotation Round With Nine Pathologists

Marker	Percent Agreement in Nine Laboratories*	Agreement in Pairs of Laboratories		κ Statistic	
		%	Bootstrap SE	Generalized	Bootstrap SE
BCL2	46	70	8	0.45	0.12
BCL2 intensity	33	74	5	0.51	0.11
BCL6 laboratory 1	12	53	7	0.42	0.07
BCL6 laboratory 7	6	54	5	0.37	0.07
BCL6 laboratory 7 (weak versus strong)	42	80	6	0.58	0.11
CD10	70	87	6	0.72	0.12
CD20†	100	100	0	—	—
CD5 laboratory 7	46	73	8	0.43	0.13
CD5 laboratory 3	49	71	8	0.44	0.12
CD5 laboratory 3 ($\leq$ 75% $v >$ 75%)	67	86	5	0.45	0.18
HLADR	88	95	4	0.75	0.13
Ki67	6	58	4	0.39	0.07
Ki67 (four categories)	61	83	6	0.58	0.11
MUM1	6	54	6	0.41	0.07
GCB	3	77	7	0.62	0.11

*Computed as the percent of the 36 patients among whom the nine scores agree.
†κ statistic cannot be computed for CD20 because 100% of the patients scored are positive, resulting in 100% expected agreement and denominator of the κ statistic equal to 0.

of great benefit, but it certainly hampers the comparison of published literature and may show unexpected results (as for CD5 in this study). In our series, maximal enhancement of bcl-6 staining showed that all DLBCL expressed bcl-6 protein to some extent. This is in line with gene expression results.[1] The reproducibility of the stainings with maximal enhancement, however, certainly was not improved as a result of arbitrary cutoff levels (bcl-6) or high percentages of patients who could not be scored due background staining (CD5).

Inability to assess a staining result in an individual patient may be an underestimated problem. Apart from the TMA-specific problem of missing cores, reasons for excluding a patient for a specific marker were different types of technical artifacts, especially high background staining and absence of an internal control. The latter aspect was encountered mostly for bcl-6 and CD10. For all other stains, admixed T cells are always present and can serve as internal control. Given that CD10 is situated at the base of the most commonly accepted algorithm to distinguish GCB- versus ABC-like DLBCL subtypes, and bcl-6 is the second dominant discriminator for GCB-like DLBCL, these effects may have consequences for class assignment in different studies. Indeed, excluding patients who could not be scored, a very high agreement across nine laboratories was seen (89% agreement; $\kappa = 0.77$) that decreased considerably when the category of patients who could not be scored was considered in the analysis (77% agreement; $\kappa = 0.62$). The fact that it is only possible to reach an acceptable level of interobserver reproducibility in this highly controlled scoring protocol is a strong argument for limiting treatment stratification on the basis of GCB versus non-GCB features to clinical trials with central pathology review support. In fact, this level of reproducibility in optimized stains is fully in line with the situation for Her2-immunostaining in breast cancer.[27] In the published series thus far, all mentioned aspects that result in scoring variation may play a role in the variation of the proportion of non–GCB- versus GCB-like subtype, varying between 65%/35%[5] versus 51%/49%.[9] Specialized series fall significantly outside these ranges, however, also suggesting a true biologic selection (ABC/GCB 68%/32% in refractory and relapsed patients,[28] 17%/83% in pediatric patients[29]).

Taken together, this study shows that semiquantitative immunohistochemistry for prognostic stratification of DLBCL is feasible in a reproducible way but with varying rates of success for different markers. Lack of harmonization of techniques and interpretation is likely to explain in part the wide variability of published results. At this stage, clinical decisions based on immunohistochemical stratification should only be performed in the context of clinical trials with centralized consensus review and validated assessment of biomarkers, and not on results of individual local centers. The LLBC will now use this approach in the first international study including a large cohort of uniformly treated DLBCL patients from collaborative clinical studies to evaluate their clinical and biologic significance.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The authors indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: Daphne de Jong, Andreas Rosenwald, Elias Campo, Anton Hagenbeek, Randy D. Gascoyne, Gilles Salles, Edie Weller
Provision of study materials or patients: Daphne de Jong, Andreas Rosenwald, Mukesh Chhanabhai, Philippe Gaulard, Wolfram Klapper, Abigail Lee, Birgitta Sander, Andrew Norton, Randy D. Gascoyne
Collection and assembly of data: Daphne de Jong, Andreas Rosenwald, Mukesh Chhanabhai, Philippe Gaulard, Wolfram Klapper, Abigail Lee, Birgitta Sander, Christoph Thorns, Elias Campo, Thierry Molina, Andrew Norton, Randy D. Gascoyne, Edie Weller
Data analysis and interpretation: Daphne de Jong, Andreas Rosenwald, Mukesh Chhanabhai, Philippe Gaulard, Wolfram Klapper, Birgitta Sander, Christoph Thorns, Elias Campo, Thierry Molina, Anton Hagenbeek, Sandra J. Horning, Andrew Lister, John Raemaekers, Randy D. Gascoyne, Gilles Salles, Edie Weller
Manuscript writing: Daphne de Jong, Andreas Rosenwald, Gilles Salles, Edie Weller
Final approval of manuscript: Daphne de Jong, Andreas Rosenwald, Mukesh Chhanabhai, Philippe Gaulard, Wolfram Klapper, Abigail Lee, Birgitta Sander, Christoph Thorns, Elias Campo, Thierry Molina, Andrew Norton, Anton Hagenbeek, Sandra Horning, Andrew Lister, John Raemaekers, Randy D. Gascoyne, Gilles Salles, Edie Weller

REFERENCES

1. Alizadeh AA, Eisen MB, Davis RE, et al: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511, 2000

2. Rosenwald A, Wright G, Chan WC, et al: The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346:1937-1947, 2002

3. Monti S, Savage KJ, Kutok JL, et al: Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood 105:1851-1861, 2005

4. Feuerhake F, Kutok JL, Monti S, et al: NFkappaB activity, function, and target-gene signatures in primary mediastinal large B-cell lymphoma and diffuse large B-cell lymphoma subtypes. Blood 106:1392-1399, 2005

5. Barrans SL, Carter I, Owen RG, et al: Germinal center phenotype and bcl-2 expression combined with International Prognostic Index improves patient risk stratification in diffuse large B-cell lymphoma. Blood 99:1136-1143, 2002

6. Colomo L, Lopez-Guillermo A, Perales M, et al: Clinical impact of the differentiation profile assessed by immunophenotyping in patients with diffuse large B-cell lymphoma. Blood 101:78-84, 2003

7. Hans CP, Weisenburger DD, Greiner TC, et al: Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. Blood 103:275-282, 2004

8. Chang CC, McClintock S, Cleveland RP, et al: Immunohistochemical expression patterns of germinal center and activation B-cell markers correlate with prognosis in diffuse large B-cell lymphoma. Am J Surg Pathol 28:464-470, 2004

9. de Paepe P, Achten R, Verhoef G, et al: Large cleaved and immunoblastic lymphoma may represent two distinct clinicopathological entities within the group of diffuse large B-cell lymphomas. J Clin Oncol 23:1-9, 2005

10. Berglund M, Thunberg U, Amini R-M, et al: Evaluation of immunophenotype in diffuse large B-cell lymphoma and its impact on prognosis. Mod Pathol 18:1113-1120, 2005

11. Zinzani PL, Dirnhofer S, Sabatini E, et al: Identification of outcome predictors in diffuse large B-cell lymphoma: Immunohistochemical profiling of homogeneously treated de novo tumors with nodal presentation on tissue microarrays. Haematologica 90:341-347, 2005

12. Kramer MH, Hermans J, Parker J, et al: Clinical significance of bcl2 and p53 protein expression in diffuse large B-cell lymphoma: A population-based study. J Clin Oncol 14:2131-2138, 1996

13. Hill ME, MacLennan KA, Cunningham DC, et al: Prognostic significance of bcl-2 expression and bcl-2 major breakpoint region rearrangement in diffuse large cell non-Hodgkin's lymphoma: A British National Lymphoma Investigation Study. Blood 88: 1046-1051, 1996

14. Hermine O, Haioun C, Lepage E, et al: Prognostic significance of bcl-2 protein expression in aggressive non-Hodgkin's lymphoma. Blood 87:265-272, 1996

15. Gascoyne RD, Adomat SA, Krajewski S, et al: Prognostic significance of bcl-2 protein expression and bcl-2 gene rearrangement in diffuse aggressive non-Hodgkin's lymphoma. Blood 90:244-251, 1997

16. Iqbal J, Sanger WG, Horsman DE, et al: Bcl-2 translocation defines a unique tumor subset within the germinal center B-cell-like diffuse large B-cell lymphoma. Am J Pathol 165:159-166, 2004

17. Lossos IS, Jones CD, Warnke R, et al: Expression of a single gene, BCL-6, strongly predicts survival in patients with diffuse large B-cell lymphoma. Blood 98:945-951, 2001

18. Adida C, Haioun C, Gaulard P, et al: Prognostic significance of survivin expression in diffuse large B-cell lymphomas. Blood 96:1921-1925, 2000

19. Sagaert X, de Paepe P, Libbrecht L, et al: Forkhead box protein P1 expression in mucosa-associated lymphoid tissue lymphomas predicts poor prognosis and transformation to diffuse large B-cell lymphoma. J Clin Oncol 24:2490-2497, 2006

20. Banham AH, Connors JM, Brown PJ, et al: Expression of the FOXP1 transcription factor is strongly associated with inferior survival in patients with diffuse large B-cell lymphoma. Clin Cancer Res 11:1065-1072, 2005

21. Barrans SL, Fenton JA, Banham A, et al: Strong expression of FOXP1 identifies a distinct subset of diffuse large B-cell lymphoma (DLBCL) patients with poor outcome. Blood 104:2933-2935, 2004

22. Kersten MJ, de Jong D, Raemaekers J, et al: Beyond the International Prognostic Index: New prognostic factors in follicular lymphoma and diffuse large-cell lymphoma—A meeting report of the Second International Lunenburg Lymphoma Workshop. Hematol J 5:202-208, 2004

23. Woolson RF, Clarke WR: Statistical Methods for the Analysis of Biomedical Data. New York, NY, Wiley, 2002

24. Efron B, Tibshirani RJ: An Introduction to the Bootstrap. New York, NY, Chapman & Hall, 1993

25. Mounier N, Briere J, Gisselbrecht 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 101:4279-4284, 2003

26. Iqbal J, Neppalli VT, Wright G, et al: Bcl-2 expression is a prognostic marker for activated B-cell-like type of diffuse large B-cell lymphoma. J Clin Oncol 24:961-968, 2006

27. Perez EA, Suman VJ, Davidson NE, et al: HER2 testing by local, central and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. J Clin Oncol 24:3032-3038, 2006

28. Moskowitz CH, Zelenetz AD, Kwealramani T, 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 106:3383-3385, 2005

29. Oschlies I, Klapper W, Zimmerman M, et al: Diffuse large B-cell lymphoma in paediatric patients predominantly belong to the germinal-center type B-cell lymphomas. Blood 107:4047-4052, 2006

Appendix

Contributors. The Lunenburg Lymphoma Biomarker Consortium is a collaboration of nine international lymphoma collaborative groups, each represented by a clinical investigator and one or more hematopathologists, and supported by a team of statisticians. EORTC Lymphoma Group: Daphne de Jong, Dennis Veldhuizen, John Raemaekers. HOVON: Daphne de Jong, Marie José Kersten, Anton Hagenbeek. GELA: Philippe Gaulard, Thierry Molina, Josette Briere, Gilles Salles. British Columbia Cancer Center: Randy Gascoyne, Mukesh Chhanabhai, Laurie Sehn. ECOG: Randy Gascoyne, Sandra Horning. German High Grade Non-Hodgkin's Lymphoma Group (DSHNHL): Christoph Thorns, Andreas Rosenwald, Wolfram Klapper, German Ott, Sylvia Hoeller, Heinz-Wolfram Bernd, Michael Pfreundschuh. NLSG: Birgitta Sander, Eva Kimby. St Bartholomew's Hospital: Abigail Lee, Andrew Norton, Andrew Clear, Andrew Lister. Independent pathology advisor: Elias Campo, Barcelona, Spain; and local support: Antoni Martinez. Dana-Farber Cancer Institute: Edie Weller.