Predictive Biomarkers and Personalized Medicine

A Comparative Analysis of Prognostic Factor Models for Follicular Lymphoma Based on a Phase III Trial of CHOP–Rituximab versus CHOP + 131Iodine—Tositumomab

Oliver W. Press1,3, Joseph M. Unger2, Lisa M. Rimsza4, Jonathan W. Friedberg6, Michael LeBlanc2, Myron S. Czuczman7,8, Mark Kaminski9, Rita M. Braziel10, Catherine Spier4, Ajay K. Gopal1,3, David G. Maloney1,3, Bruce D. Cheson8,11, Shaker R. Dakhil12, Thomas P. Miller5, and Richard I. Fisher13

Abstract

Purpose: There is currently no consensus on optimal frontline therapy for patients with follicular lymphoma. We analyzed a phase III randomized intergroup trial comparing six cycles of CHOP-R (cyclophosphamide–Adriamycin–vincristine–prednisone (Oncovin)–rituximab) with six cycles of CHOP followed by iodine-131 tositumomab radioimmunotherapy (RIT) to assess whether any subsets benefited more from one treatment or the other, and to compare three prognostic models.

Experimental Design: We conducted univariate and multivariate Cox regression analyses of 532 patients enrolled on this trial and compared the prognostic value of the FLIPI (follicular lymphoma international prognostic index), FLIPI2, and LDH + β2M (lactate dehydrogenase + β2-microglobulin) models.

Results: Outcomes were excellent, but not statistically different between the two study arms [5-year progression-free survival (PFS) of 60% with CHOP-R and 66% with CHOP-RIT (P = 0.11); 5-year overall survival (OS) of 92% with CHOP-R and 86% with CHOP-RIT (P = 0.08); overall response rate of 84% for both arms]. The only factor found to potentially predict the impact of treatment was serum β2M; among patients with normal β2M, CHOP-RIT patients had better PFS compared with CHOP-R patients, whereas among patients with high serum β2M, PFS by arm was similar (interaction P value = 0.02).

Conclusions: All three prognostic models (FLIPI, FLIPI2, and LDH + β2M) predicted both PFS and OS well, though the LDH + β2M model is easiest to apply and identified an especially poor risk subset. In an exploratory analysis using the latter model, there was a statistically significant trend suggesting that low-risk patients had superior observed PFS if treated with CHOP-RIT, whereas high-risk patients had a better PFS with CHOP-R. Clin Cancer Res; 19(23); 6624–32. ©2013 AACR.

Introduction

Follicular lymphoma is a common, indolent non-Hodgkin lymphoma (NHL) associated with long-term survival with a variety of initial treatment approaches (1, 2). Recent longitudinal and epidemiologic studies suggest that survival of follicular lymphoma patients has markedly improved in the last 15 years, concurrent with the implementation of immunochemotherapy regimens incorporating both chemotherapy and anti-CD20 monoclonal antibodies (3–8), but there is no consensus on which of these regimens is optimal. In an attempt to address this question, SWOG and Cancer and Leukemia Group B (CALGB) designed a phase III study in 1999 to 2000 comparing two of the most promising chemotherapy regimens for follicular lymphoma at the time, namely, 6 cycles of CHOP [cyclophosphamide–Adriamycin–vincristine–prednisone (Oncovin)] chemotherapy administered with 6 doses of rituximab versus 6 cycles of CHOP chemotherapy, followed by dosimetric and therapeutic doses of tositumomab and 131I-tositumomab as consolidative radioimmunotherapy (RIT), based on previous promising pilot studies of these regimens (9–11). The results of this phase III trial (S0016) have recently been reported (12) and demonstrated that the progression-free survival (PFS) and overall survival (OS) were excellent on both arms of the study, but not statistically

Authors' Affiliations: 1Clinical Research Division; 2SWOG Statistical Center; 3Fred Hutchinson Cancer Research Center; 4Medical Oncology, University of Washington Medical Center, Seattle, Washington; 5Department of Pathology; 6Arizona Cancer Center, University of Arizona, Tucson, Arizona; 7James P. Wilmot Cancer Center, University of Rochester Medical Center, Rochester; 8Roswell Park Cancer Institute, Buffalo, New York; 9Cancer and Leukemia Group B; 10University of Michigan, Ann Arbor, Michigan; 11Department of Pathology, Oregon Health and Science University, Portland, Oregon; 12Georgetown University Hospital, Washington; 13Wichita Community Clinical Oncology Program, Wichita, Kansas; and 14Fox Chase Cancer Center—Temple Health, Temple University, Philadelphia PA.

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

O.W. Press and J.M. Unger contributed equally to this work.

Corresponding Author: Oliver W. Press, Fred Hutchinson Cancer Research Center, D3-190, 1100 Fairview Avenue, Seattle, WA 98109.

Phone: 206-667-1872; Fax: 206-667-1874; E-mail: press@u.washington.edu

doi: 10.1158/1078-0432.CCR-13-1120

©2013 American Association for Cancer Research.
These exploratory analyses. LDH (lactate dehydrogenase) and model consisting of only the baseline lactate dehydrogenase (LDH) and assesses their relative merits. It presents a Cox analysis suggesting that patients with a normal β2-microglobulin (β2M) level treated with CHOP-RIT (radioimmunotherapy) show an improved progression-free survival (PFS) compared with those with CHOP-R (76% vs. 61% after 5 years, respectively; interaction P value = 0.02). It also compares three prognostic models for follicular lymphoma [(FLIPI1, FLIPI2, and β2M + LDH (lactate dehydrogenase)] and assesses their relative value.

**Translational Relevance**

This article reports results from the largest phase III clinical trial of follicular lymphoma conducted in North America in the last 20 years, comparing two immunotherapy options, namely, CHOP-R (cyclophosphamide–Adriamycin–vincristine–prednisone (Oncovin)–rituximab) and CHOP + 131Iodine–tositumomab. It presents a Cox analysis suggesting that patients with a normal β2-microglobulin (β2M) level treated with CHOP-RIT (radioimmunotherapy) show an improved progression-free survival (PFS) compared with those with CHOP-R (76% vs. 61% after 5 years, respectively; interaction P value = 0.02). It also compares three prognostic models for follicular lymphoma [(FLIPI1, FLIPI2, and β2M + LDH (lactate dehydrogenase)] and assesses their relative merits. The findings of this large study will be helpful to hematologists and oncologists as they evaluate their patients and choose immunotherapy regimens.

Different with 4.9 years of median follow-up. It remains possible, however, that some subsets of patients might benefit more from one regimen or the other. To address this hypothesis, we conducted an exploratory analysis using univariate and multivariate Cox regression to identify subgroups of patients with follicular lymphoma with differential outcomes using CHOP–rituximab (CHOP-R) or CHOP-RIT. In addition, we used this dataset to compare and contrast the relative values of three prognostic models for follicular lymphoma, namely, the original follicular lymphoma international prognostic index (FLIPI) model (13), an updated FLIPI2 model (14), or a laboratory-based model consisting of only the baseline lactate dehydrogenase (LDH) and β2M values. This article presents the results of these exploratory analyses.

**Materials and Methods**

**Eligibility**

Details of the protocol eligibility and exclusion criteria have been published elsewhere (12). In brief, patients older than 18 years with untreated, measurable bulky stage II or stage III to IV follicular lymphoma (grade 1, 2, or 3) expressing CD20 were eligible if they had a SWOG performance status of 0 to 2, granulocytes ≥ 1,500 cells/μL, and platelets ≥ 100,000/μL. Bulky adenopathy was defined as more than 10 cm in diameter or greater than one third the thoracic diameter. Excisional biopsies or large core needle biopsies showing follicular architecture were required; fine needle aspirates and marrow biopsies alone were not sufficient. Diagnostic biopsies were all reviewed centrally by expert SWOG pathologists to confirm the diagnosis of follicular lymphoma according to published consensus morphologic, immunophenotypic, and genetic criteria (15). Cases with more than 25% diffuse architecture and more than 15 centroblasts per high power field were considered diffuse large B cell lymphoma and excluded. Investigators were asked to enroll only patients with follicular lymphoma requiring therapy and not asymptomatic, low tumor burden patients for whom watchful waiting would be appropriate. All patients signed a written informed consent in accordance with institutional and federal guidelines.

**Study design and protocol treatment**

Baseline and serial follow-up patient evaluations, laboratory testing, and imaging studies were performed as previously described (12). Patients received CHOP chemotherapy every 21 days for 6 cycles using standard doses, supportive care, and dose reductions as published (10–12). Patients on the CHOP-R arm were treated as described by Czuczman and colleagues with 375 mg/m² of rituximab on days 1, 6, 48, 90, 134, and 141 and CHOP chemotherapy on days 8, 29, 50, 71, 92, and 113 (16). Patients on the CHOP-RIT arm received consolidative RIT with tositumomab/131I–tositumomab administrated 4 to 8 weeks after the sixth cycle of CHOP (10–12).

**Statistical analysis**

The primary objective of the research trial was to determine which of the two regimens tested (CHOP-R or CHOP-RIT) was superior in terms of PFS. Patients were randomized to CHOP-R or CHOP-RIT using a dynamic allocation scheme at the time of registration. Data were centrally reviewed and clinical responses assigned by the SWOG statistical center and the principal investigator according to the criteria established in two international workshops (17, 18). Remission and survival status was assessed 200 days and 365 days after initiation of therapy and then every 6 months until death. Restaging was also performed whenever patients developed symptoms or signs of relapse. Patient randomization was balanced according to β2-microglobulin (β2M) level >1ULN (institutional upper limit of normal) vs. ≤1ULN. PFS was defined as the time from registration to the first observation of progressive disease or death due to any cause. Analyses of PFS and OS were performed on the basis of modified “intent-to-treat,” in which only patients known to be ineligible were excluded. Multivariate PFS and OS analyses were performed by Cox regression (19) and survival was estimated according to the method of Kaplan–Meier (20). This study was continuously monitored by a Data and Safety Monitoring Committee and formal interim analyses were performed after 50% and after 75% of eligible patients were randomized.

**Prognostic factor analyses**

Univariate analyses of the association of baseline clinical and selected laboratory factors with both PFS and OS were performed, with adjustment for treatment arm. In addition, we assessed whether the association of any factor and PFS or OS differed according to treatment using interaction terms in multivariate Cox regression models. Because the analyses were exploratory, the results were not corrected for multiple comparisons.
**Prognostic model comparison**

FLIPI and FLIPI2 risk scores were calculated as described in the original publications (13, 14). The chart of each patient was individually reviewed by both the principal investigator and by a data coordinator in the SWOG statistical center to confirm accurate assignment of patients to risk groups according to these models. In particular, careful attention was paid to lymph node group assignments with comparison of baseline tumor assessment forms and radiographic reports to published FLIPI lymph node maps (13). In addition, maximal lymph node diameters were double checked for each patient to ensure correct designation of FLIPI2 scores. We compared the performance of the FLIPI and FLIPI2 models with a model based entirely on baseline laboratory tests. Candidate factors for this laboratory-based risk model must have been significant predictors of both PFS and OS in both the univariate and multivariate regression settings. The factors that remained statistically significant independent predictors of both PFS and OS were standardized according to institutional normal limits. A panel of risk models was then generated on the basis of varying the split points of the continuous factors together in increments of 10%. Best laboratory-based risk models were determined by comparing Wald $\chi^2$ statistics. Models were fitted using both ordinal categorical variables and dummy variables. To adjust for multiple comparisons permutation sampling was used to control the family-wise type I error for each cutoff point, based on 1,000 samples (21).

**Results**

**Patient characteristics**

Key baseline characteristics of the 532 eligible patients enrolled on this study are as follows: median age, 54 years (range: 23–87 years); male gender, 54%; white race, 90%; elevated $\beta$2M, 54%; median $\beta$2M, 2.2 (range: 0.1–41); B symptoms, 27%; bulk more than 10 cm or more than one-third of the thoracic diameter, 25%; stage IV, 61%; and intermediate or high-risk FLIPI scores, 70%. These (and other) baseline characteristics were well balanced between the two treatment arms without any statistically significant differences (12). Only 3% of patients had missing baseline LDH, with no difference by arm; all patients had baseline $\beta$2M data.

**Clinical outcomes**

After a median follow-up of 4.9 years, the estimated 5-year PFS was 66% with CHOP-RIT and 60% with CHOP-R (two-sided $P = 0.11$) and estimates of 5-year OS were 86% with CHOP-RIT and 92% with CHOP-R ($P = 0.08$). The overall response rate was 84% in both the CHOP-R and CHOP-RIT arms, whereas the complete remission (CR) rate was (40%) in the CHOP-R arm and (45%) in the CHOP-RIT arm ($P = 0.30$). We further analyzed survival outcomes according to type of remission, though to avoid potential selection and lead-time biases, we only considered outcomes in responders who had achieved at least 1-year survival without progression using landmark survival analysis. The 5-year PFS for patients with partial remissions (PR) was 61%, compared with 70% for patients with CRs ($P = 0.01$). There was no difference in 5-year OS between patients with partial (91%) and complete (93%) remissions ($P = 0.28$). Finally, there was no evidence that the pattern of better PFS for CR patients and similar OS between CR and PR patients were different within treatment arms (data not shown).

**Prognostic factor analysis**

Univariate and multivariate Cox proportional hazards regression analyses were performed to identify clinical features significantly associated with PFS and OS (19). Table 1 shows univariate results by baseline factors. Seven factors were significant predictors of PFS at the two-sided, $\alpha = 0.05$ level, including maximal lymph node diameter, hemoglobin, performance status, $\beta$2M, LDH, number of lymph nodes, and stage (listed in order of decreasing statistical significance). A statistically significant interaction with treatment was evident in only one case; specifically the association between treatment and PFS differed by $\beta$2M levels (interaction $P$ value = 0.02; Fig. 1). In patients with a normal baseline $\beta$2M level, those treated with CHOP-RIT had a better PFS than those treated with CHOP-R (76% vs. 61% after 5 years), whereas patients with an elevated $\beta$2M had similar PFS with both regimens. Six factors were statistically significantly associated with OS in a univariate Cox regression analysis, including maximum lymph node diameter, performance status, LDH, $\beta$2M, age, and hemoglobin (listed in order of decreasing statistical significance; Table 1). None of these variables, however, interacted significantly with treatment arm at the $P = 0.05$ level.

We conducted a multivariable analysis of the factors that were statistically significantly associated with both PFS and OS (i.e., hemoglobin, LDH, maximal lymph node size, performance status, and $\beta$2M). The results showed that hemoglobin (HR, 1.96; 95% confidence interval (CI), 1.29–2.97; $P = 0.002$) and maximal lymph node size (HR, 1.81; 95% CI, 1.31–2.49; $P = 0.0003$) retained statistical significance in the multivariable model for PFS, and that LDH (HR, 1.90; 95% CI, 1.09–3.31; $P = 0.02$) and maximal lymph node size (HR, 1.96; 95% CI, 1.10–3.49; $P = 0.02$) retained statistical significance in a multivariable model for OS (Supplementary Table S1).

**Comparison of three prognostic factor models for follicular lymphoma**

We next conducted a multivariable Cox regression analysis to test the ability of the FLIPI and FLIPI2 prognostic models to predict outcomes of patients treated on this protocol (Table 2). For the FLIPI model, each increase in risk level was associated with a HR of 1.55 for PFS ($P < 0.0001$) and 1.95 for OS ($P = 0.001$). The FLIPI2 model performed even better, with each increase in risk level associated with a HR of 1.90 for PFS ($P < 0.0001$) and 2.65 for OS ($P < 0.0001$). Figure 2 shows PFS and OS
Kaplan–Meier curves for each model. There was marginal evidence that the association of FLIPI and PFS differed by treatment arm (interaction \( P \) value = 0.08). A comparison of the risk group distributions in our trial with those of selected other major follicular lymphoma studies demonstrates that our study had fewer low-risk patients than either the original FLIPI (13) or FLIPI2 studies (14), but somewhat more than the recent Primary Rituximab and Maintenance (PRIMA) trial (ref. 6; Supplementary Table S2).

We also derived a prognostic model based exclusively on routine laboratory tests. Univariate models demonstrated that \( \beta2M \), LDH, and hemoglobin were all statistically significantly associated with PFS and OS (Table 1), but only \( \beta2M \) and LDH remained independent for both PFS and OS in multivariate models that included all three factors (\( P \leq 0.05 \)). Therefore, among the three laboratory variables analyzed, the candidate laboratory-based factors for our model were \( \beta2M \) and LDH. Both factors were standardized, by dividing the observed value by the IULN, creating a ratio. Each ratio, in parallel, was then allowed to vary by a percentage of the IULN. For instance, patients were categorized according to whether both factors were less than 150% of IULN (coded 0), one or the other factor was greater than 150% of IULN (coded 1), or both factors were greater than 150% of IULN (coded 2). A range of models with ratios from 50% of IULN to 200% (twice) the IULN was explored in 10% increments on each variable; models were then compared using model Wald \( \chi^2 \) statistics. As each model had zero, one, or two risk factors, we explored models using both ordinal categorical variables, which assumes a linear

**Table 1.** Univariate analyses of baseline factors of 532 eligible patients enrolled on intergroup protocol S0016

<table>
<thead>
<tr>
<th>Factor</th>
<th>Factor levels (codes)</th>
<th>CHOP-RIT</th>
<th>CHOP-R</th>
<th>PFS Univariate HR (95% CI)</th>
<th>Interaction with therapy</th>
<th>OS Univariate HR (95% CI)</th>
<th>Interaction with therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>≤60 (0)</td>
<td>72%</td>
<td>69%</td>
<td>1.21 (0.83–1.52)</td>
<td>( P = 0.10 )</td>
<td>1.94 (1.18–3.18)</td>
<td>( P = 0.06 )</td>
</tr>
<tr>
<td></td>
<td>&gt;60 (1)</td>
<td>28%</td>
<td>31%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bone Marrow involvement</td>
<td>No (0)</td>
<td>44%</td>
<td>44%</td>
<td>1.29 (0.96–1.72)</td>
<td>( P = 0.73 )</td>
<td>1.00 (0.61–1.62)</td>
<td>( P = 0.35 )</td>
</tr>
<tr>
<td></td>
<td>Yes (1)</td>
<td>56%</td>
<td>56%</td>
<td></td>
<td>( P = 0.09 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>≥12 (0)</td>
<td>89%</td>
<td>89%</td>
<td>2.14 (1.46–3.16)</td>
<td>( P = 0.0001 )</td>
<td>2.14 (1.13–4.05)</td>
<td>( P = 0.39 )</td>
</tr>
<tr>
<td></td>
<td>&lt;12 (1)</td>
<td>11%</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDH</td>
<td>≤IULN (0)</td>
<td>74%</td>
<td>78%</td>
<td>1.60 (1.17–2.18)</td>
<td>( P = 0.59 )</td>
<td>2.45 (1.48–4.05)</td>
<td>( P = 0.79 )</td>
</tr>
<tr>
<td></td>
<td>&gt;IULN (1)</td>
<td>26%</td>
<td>22%</td>
<td></td>
<td>( P = 0.0003 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximal LN size</td>
<td>≤6 cm (0)</td>
<td>61%</td>
<td>60%</td>
<td>1.95 (1.47–2.59)</td>
<td>( P = 0.20 )</td>
<td>2.62 (1.59–4.30)</td>
<td>( P = 0.34 )</td>
</tr>
<tr>
<td></td>
<td>&gt;6 cm (1)</td>
<td>39%</td>
<td>40%</td>
<td></td>
<td>( P &lt; 0.0001 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of nodes</td>
<td>≤4 (0)</td>
<td>61%</td>
<td>60%</td>
<td>1.52 (1.14–2.02)</td>
<td>( P = 0.75 )</td>
<td>2.22 (0.15–9.99)</td>
<td>( P = 0.92 )</td>
</tr>
<tr>
<td></td>
<td>&gt;4 (1)</td>
<td>39%</td>
<td>40%</td>
<td></td>
<td>( P = 0.004 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance status</td>
<td>0 (0)</td>
<td>74%</td>
<td>74%</td>
<td>1.72 (1.27–2.34)</td>
<td>( P = 0.93 )</td>
<td>2.12 (1.27–3.51)</td>
<td>( P = 0.69 )</td>
</tr>
<tr>
<td></td>
<td>1 or 2 (1)</td>
<td>26%</td>
<td>26%</td>
<td></td>
<td>( P = 0.0004 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Female (0)</td>
<td>45%</td>
<td>47%</td>
<td>1.12 (0.84–1.49)</td>
<td>( P = 0.68 )</td>
<td>1.50 (0.90–2.47)</td>
<td>( P = 0.30 )</td>
</tr>
<tr>
<td></td>
<td>Male (1)</td>
<td>55%</td>
<td>53%</td>
<td></td>
<td>( P = 0.43 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serum ( \beta2M )</td>
<td>≤IULN (0)</td>
<td>46%</td>
<td>47%</td>
<td>1.69 (1.26–2.27)</td>
<td>( P = 0.02 )</td>
<td>2.22 (1.31–3.75)</td>
<td>( P = 0.10 )</td>
</tr>
<tr>
<td></td>
<td>&gt;IULN (1)</td>
<td>54%</td>
<td>53%</td>
<td></td>
<td>( P = 0.0004 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spleen involvement</td>
<td>No (0)</td>
<td>77%</td>
<td>80%</td>
<td>1.14 (0.81–1.59)</td>
<td>( P = 0.06 )</td>
<td>0.85 (0.46–1.56)</td>
<td>( P = 0.10 )</td>
</tr>
<tr>
<td></td>
<td>Yes (1)</td>
<td>23%</td>
<td>20%</td>
<td></td>
<td>( P = 0.45 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td>&lt;Stage IV (0)</td>
<td>36%</td>
<td>41%</td>
<td>1.42 (1.05–1.93)</td>
<td>( P = 0.19 )</td>
<td>1.11 (0.67–1.85)</td>
<td>( P = 0.49 )</td>
</tr>
<tr>
<td></td>
<td>IV (1)</td>
<td>64%</td>
<td>59%</td>
<td></td>
<td>( P = 0.02 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td>A (0)</td>
<td>74%</td>
<td>71%</td>
<td>0.85 (0.61–1.17)</td>
<td>( P = 0.19 )</td>
<td>1.16 (0.68–1.96)</td>
<td>( P = 0.65 )</td>
</tr>
<tr>
<td></td>
<td>B (1)</td>
<td>26%</td>
<td>29%</td>
<td></td>
<td>( P = 0.32 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Figure 1. PFS of advanced follicular lymphoma patients enrolled on intergroup protocol S0016, stratified according to treatment arm and by serum \( \beta2M \) level. Five year Kaplan-Meier estimates are shown.](image-url)
association—that is, the same increase in risk moving from low to intermediate risk as from intermediate to high risk—as well as a dummy variable model, which does not make a linearity assumption. Supplementary Fig. S1 shows how the Wald $\chi^2$ values, for both PFS (top) and OS (bottom), achieved maxima at 150% of the IULN. The close tracking of the respective curves for the dummy variable models and the ordinal categorical models suggest an approximately linear association between low, intermediate, and high categories, so the ordinal categorical approach was chosen for analysis. Importantly, multiple comparisons testing showed that each of the cutoff points at 70% of IULN or above was statistically significant at the permutation-adjusted $P = 0.01$ level.

Two models were constructed, one splitting each factor at the IULN (for clinical simplicity) and the other model at the optimal cutoff point of 150% of the IULN (for maximum separation of prognostic groups). Results are shown in the bottom of Table 2 and in Fig. 3. For the model splitting at the IULN, each increase in risk level was associated with a HR of 1.61 for PFS ($P < 0.0001$; Fig. 3A) and 2.05 for OS ($P < 0.0001$; Fig. 3B). For the model splitting at 150% of IULN, each increase in risk level was associated with a HR of 2.00 for PFS ($P < 0.0001$; Fig. 3C) and 2.61 for OS ($P < 0.0001$; Fig. 3D).

**Table 2.** Comparison of multivariable models for outcome parameters of patients enrolled on intergroup protocol S0016

<table>
<thead>
<tr>
<th>Factor</th>
<th>% in Group (codes)</th>
<th>Overall</th>
<th>CHOP-RIT</th>
<th>CHOP-R</th>
<th>PFS Univariate HR (95% CI)</th>
<th>PFS Interaction with Tx</th>
<th>OS Univariate HR (95% CI)</th>
<th>OS Interaction with Tx</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLIPI Low (0)</td>
<td>30%</td>
<td>31%</td>
<td>30%</td>
<td></td>
<td>1.55 (1.27–1.88)</td>
<td>0.08</td>
<td>1.95 (1.38–2.73)</td>
<td>0.27</td>
</tr>
<tr>
<td>FLIPI Intermediate (1)</td>
<td>45%</td>
<td>43%</td>
<td>47%</td>
<td></td>
<td>(1.27–1.88)</td>
<td>$P &lt; 0.0001$</td>
<td>(1.38–2.73)</td>
<td>$P = 0.0001$</td>
</tr>
<tr>
<td>FLIPI High (2)</td>
<td>24%</td>
<td>26%</td>
<td>22%</td>
<td></td>
<td>1.90 (1.49–2.41)</td>
<td>0.11</td>
<td>2.65 (1.72–4.07)</td>
<td>0.11</td>
</tr>
<tr>
<td>FLIPI Low (0)</td>
<td>11%</td>
<td>13%</td>
<td>10%</td>
<td></td>
<td>(1.49–2.41)</td>
<td>$P &lt; 0.0001$</td>
<td>(1.72–4.07)</td>
<td>$P &lt; 0.0001$</td>
</tr>
<tr>
<td>FLIPI Intermediate (1)</td>
<td>59%</td>
<td>56%</td>
<td>62%</td>
<td></td>
<td>1.61 (1.33–1.96)</td>
<td>0.03</td>
<td>2.05 (1.46–2.87)</td>
<td>0.23</td>
</tr>
<tr>
<td>FLIPI High (2)</td>
<td>30%</td>
<td>31%</td>
<td>28%</td>
<td></td>
<td>(1.33–1.96)</td>
<td>$P &lt; 0.0001$</td>
<td>(1.46–2.87)</td>
<td>$P &lt; 0.0001$</td>
</tr>
<tr>
<td>b2M–LDH, both split at IULN</td>
<td>Low (0)</td>
<td>41%</td>
<td>40%</td>
<td>43%</td>
<td>1.61 (1.39–2.53)</td>
<td>0.03</td>
<td>2.00 (1.81–3.75)</td>
<td>0.11</td>
</tr>
<tr>
<td>b2M–LDH, both split at IULN</td>
<td>Intermediate (1)</td>
<td>42%</td>
<td>41%</td>
<td>43%</td>
<td>(1.39–2.53)</td>
<td>$P &lt; 0.0001$</td>
<td>(1.81–3.75)</td>
<td>$P &lt; 0.0001$</td>
</tr>
<tr>
<td>b2M–LDH, both split at IULN</td>
<td>High (2)</td>
<td>17%</td>
<td>19%</td>
<td>15%</td>
<td>2.00 (1.59–2.53)</td>
<td>$P &lt; 0.0001$</td>
<td>2.61 (1.81–3.75)</td>
<td>$P &lt; 0.0001$</td>
</tr>
<tr>
<td>b2M–LDH, both split at IULN</td>
<td>Low (0)</td>
<td>75%</td>
<td>73%</td>
<td>78%</td>
<td>1.59 (1.27–1.98)</td>
<td>$P &lt; 0.0001$</td>
<td>2.65 (1.99–3.37)</td>
<td>$P &lt; 0.0001$</td>
</tr>
<tr>
<td>b2M–LDH, both split at IULN</td>
<td>Intermediate (1)</td>
<td>21%</td>
<td>23%</td>
<td>19%</td>
<td>(1.27–1.98)</td>
<td>$P &lt; 0.0001$</td>
<td>(2.65–3.37)</td>
<td>$P &lt; 0.0001$</td>
</tr>
<tr>
<td>b2M–LDH, both split at IULN</td>
<td>High (2)</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
<td>1.59 (1.27–1.98)</td>
<td>$P &lt; 0.0001$</td>
<td>2.65 (1.99–3.37)</td>
<td>$P &lt; 0.0001$</td>
</tr>
</tbody>
</table>

**Figure 2.** PFS (A and C) and OS (B and D) of advanced follicular lymphoma patients enrolled on intergroup protocol S0016, stratified according to FLIPI (A and B) or FLIPI2 risk groups (C and D). Five year Kaplan-Meier estimates are shown.
Prognostic Factor Analysis for Follicular Lymphoma

Figure 3. PFS (A and C) and OS (B and D) of advanced follicular lymphoma patients enrolled on intergroup protocol S0016, stratified according to the β2M + LDH prognostic model using either the IULN (A and B) or 150% of the IULN (C and D) as cutoff points for both variables. Five year Kaplan-Meier estimates are shown.

Fig. 3D). Figure 3C and D compare the PFS and OS, respectively, of patients in whom both LDH and β2M were above 150% of the IULN (high risk) compared with those with only one factor more than 150% of the IULN (intermediate risk), and to those with neither factor more than 150% of the IULN (low risk), and shows the outstanding separation of the curves with indicated 5-year PFS of 67%, 56%, and 20%. As illustrated in this figure, the model splitting at 150% of the IULN identifies an extremely high risk, albeit small, group of patients with very poor PFS and OS.

For both β2M + LDH models, significant interactions (P = 0.03 for each) were evident for the association of risk level and PFS, likely reflecting the influence of β2M. Low-risk patients seem to have a better PFS if treated with CHOP-RIT and high-risk patients may have a better PFS if treated with CHOP-R (Fig. 4A). High-risk patients treated with CHOP-RIT had very poor OS (Fig. 4B).

Table 2 demonstrates that in this dataset, the simple laboratory-based model split at the IULN performed as well as the FLIPI model, and the 150% model performed as well as the FLIPI 2 model and better than the original FLIPI model. The statistically significant interaction of the laboratory-based model with treatment for PFS suggests the model has predictive as well as prognostic utility. However, it should be noted that as the event rate for OS in this single phase III study was low, a split sample validation into training and test datasets was not possible. Because the 150% models represent an optimized model, confirmation that this model outperforms a simpler model based on splitting at IULN would require an independent validation dataset.

Discussion

Management strategies for follicular lymphoma remain highly controversial. There is no consensus about the optimal induction strategy for frontline therapy of indolent NHL and widely diverging views persist on the roles of rituximab “maintenance” therapy or consolidation therapy with RIT, though several well-designed phase III clinical trials have recently been reported addressing these issues. These recent studies have shown that R-CHOP and R- fludarabine-based induction regimens provide superior...
PS compared with rituximab, cyclophosphamide, vincristine, and prednisone (R-CVP), but that OS does not differ significantly with any of the regimens (6, 22). These trials also demonstrate that fludarabine-based induction regimens are more myelosuppressive and increase the risk of secondary malignancies compared with the alkylator-based regimens, which most investigators now prefer (6, 22). A single phase III study suggests that bendamustine plus rituximab induction may afford a superior PSF compared with R-CHOP, and be less toxic, though OS is similar (23). Administration of an extended course of “maintenance” rituximab for 2 years following completion of induction immunochemotherapy, or a single dose of 90Yttrium–ibritumomab–tiuxetan as consolidation have each been shown to markedly prolong PSF in patients with follicular lymphoma in separate phase III studies (6, 24, 25), but neither approach has yet been shown to improve OS compared with patients treated with induction regimens alone.

The results reported in this article are based on the largest phase III clinical trial of advanced follicular lymphoma conducted in North America in the last 20 years, which sought to assess whether follicular lymphoma patients consolidated with Iodine-131–tositumomab following 6 cycles of CHOP chemotherapy (without any rituximab exposure) would have superior outcomes compared with patients treated with 6 cycles of CHOP plus 6 doses of rituximab. The results of the trial did not demonstrate a statistically significant improvement in PSF or OS for the experimental arm (CHOP + RIT) compared with patients treated with CHOP-R, though outcomes were excellent with either regimen. Although the study demonstrated no differences between the treatment arms in the primary analysis of the entire group of patients (12), many investigators have queried whether specific subsets of patients might have benefited more from one regimen or another. We have conducted an exploratory “hypothesis-generating” subset analysis using Cox analysis to address this question, as reported in this article. In addition, we have used this large patient cohort to compare three prognostic models of follicular lymphoma to ascertain their relative merits.

Univariate regression analysis demonstrated that elevations of baseline LDH or β2M levels, anemia, large lymph node size, and poor performance status all adversely affected both PSF and OS for patients enrolled on the trial. In addition, the number of enlarged lymph nodes and the stage of disease had univariate prognostic significance for PSF but not OS. Conversely, older age was significantly associated with worse OS, but not PSF. Of all these individual factors, only an elevated β2M was shown to exhibit a statistically significant interaction with the treatment arm in terms of outcome (Table 1, Fig. 2). Patients with a normal baseline β2M level treated with CHOP-RIT showed improved PSF compared with those treated with CHOP-R (76% vs. 61% after 5 years), whereas PSF by arm in patients with high β2M was similar; (interaction P value = 0.02). This finding is consistent with several prior studies demonstrating the powerful prognostic value of β2M for patients with indolent NHL (14, 26–29), myeloma (30, 31), and other hematopoietic malignancies (32–35). Although the reason that low-risk patients might have superior PSF if treated with frontline RIT consolidation is unclear, this finding is consistent with an independent report that follicular lymphoma patients treated with 90Yttrium–ibritumomab–tiuxetan as frontline therapy had a much longer PSF if the baseline LDH was normal than if it was elevated (36). We are, of course, acutely aware of the limitations of such retrospective subset analyses, particularly when many variables are interrogated, and therefore believe that this finding should be considered hypothesis-generating rather than treatment-defining, until independent prospective studies supporting this analysis are reported.

Because the outcome of patients with follicular lymphoma is highly variable, several prognostic factor models have been developed on the basis of clinical and laboratory features to estimate the survival of newly presenting patients. These prognostic factor models have proven very useful as guides to assist patient counseling, treatment planning, and clinical trial interpretation. The most widely used model, the FLIPI, was based on a large retrospective analysis of patients diagnosed between 1985 and 1992 who were treated in the “pre-rituximab” era. This index incorporates 5 adverse prognostic factors (age >60; advanced stage; hemoglobin level <12 g/dL; >4 nodal areas of involvement; and elevated LDH) and stratifies patients into low-, intermediate-, and high-risk groups based on the presence of 0 to 1, 2, or 3 or more adverse risk features, respectively. Importantly, the original FLIPI study excluded β2M from the multivariable analysis because of the “very high proportion of patients with missing data” in the retrospective analysis (13). The FLIPI was found to be more discriminant in terms of OS than the IPI (devised on the basis of features of patients with aggressive lymphomas) and identified three groups of patients with projected 10-year OSs of 71% (low risk), 51% (intermediate risk), and 37% (high risk). The FLIPI has been widely adopted in clinical practice for prognostication and by investigators for patient stratification in clinical trials. Nevertheless, concerns have emerged about the applicability of the index in the modern era with the routine incorporation of rituximab into both frontline and salvage therapies. In addition, reservations have been expressed about the difficulties experienced by clinicians in properly assigning lymph node “groups” defined by the FLIPI model. The FLIPI lymph node “groups” do not correspond to the nodal groups defined by the Ann Arbor staging system. Furthermore, the inconsistent designation of bilateral adenopathy as constituting either one or two nodal groups has led to many errors in assigning FLIPI scores, and in the experience of SWOG, has generally led to an overestimation of the number of lymph node groups and, consequently, higher FLIPI scores. We have therefore found it essential to perform central review of all FLIPI score assignments on SWOG studies of follicular lymphoma to ensure they are accurate.

To address these and other concerns, the FLIPI investigators conducted a prospective study to confirm the value of the original FLIPI in the “rituximab era,” and also to develop
a “more accurate” prognostic index. This effort culminated in the development of the FLIPI2 based on the β2M level, the longest diameter of the largest involved lymph node (>6 cm), presence of bone marrow involvement, hemoglobin level less than 12 g/dL, and age more than 60 years. The 3-year PFS rates in the FLIPI2 study were 91%, 69%, and 51%, respectively, for patients in the low- (0–1 factors), intermediate- (2 factors), and high-risk (≥3 factors) groups. Despite its obvious strengths compared with the original FLIPI, the FLIPI2 has not been widely adopted, at least in North America, where the original FLIPI is still more widely used in both clinical practice and research trials.

We exploited the large size and prospective nature of the S0016 study to compare the strengths of the FLIPI and FLIPI2 indices with each other and with a third, simpler prognostic index based strictly on two baseline laboratory tests (LDH and β2M). Our data strongly confirm the powerful prognostic value of the FLIPI and FLIPI2 indices for stratifying patients into risk groups for both PFS and OS, and also support the contention that the FLIPI2 is superior to the original FLIPI score in separating risk groups (Fig. 1 and Table 2). Furthermore, our data suggest that the simpler LDH + β2M index is at least comparable with the FLIPI in prognostic power and may even perform as well as the FLIPI2 in defining risk groups (Table 2, Figs. 1 and 3). Importantly, this simpler model is much easier to compute and utilize. These findings strongly support a previous report from a smaller study of follicular lymphoma patients treated with fludarabine and mitoxantrone conducted by SWOG (29) and a study of intermediate grade lymphomas at The University of Texas MD Anderson Cancer Center (Houston, TX; ref. 37), validating the prognostic power of the LDH + β2M risk stratification for NHL. One of the largest unmet clinical needs in management of follicular lymphoma is the upfront identification of the small subgroup of patients destined for early relapse within 2 years of R-CHOP induction. The LDH + β2M index, particularly with the 150% cutoff point, performs exceptionally well at identifying this poor risk group, which merits consideration of frontline novel therapeutics. The results for the LDH + β2M models at cutoff points of 70% or above are statistically significant under multiple comparisons testing, however, there is less assurance that the 150% model would be the optimal model in different settings. We believe that widespread adoption of this simple, powerful prognostic index for follicular lymphoma will prove to be as useful as the International Staging System for myeloma (composed of β2M + albumin; 38) and encourage utilization of the LDH + β2M index by other investigators and cooperative groups studying follicular lymphoma. Although in this dataset, Wald χ² analysis has demonstrated the maximal power of this model using cutoff points of 150% of the IULN for both variables the model has powerful prognostic significance using any cutoff point at 70% or more of the IULN, and for simplicity in clinical practice we advocate merely using the IULNs for both variables (Fig. 3).

**Disclosure of Potential Conflicts of Interest**

O.W. Press has received commercial research support and is a consultant/advisory board member of Roche/Genentech. J.W. Friedberg and M.S. Czuczman are consultant/advisory board members of Genentech. M. Kaminski has ownership interest (including patents) in Patents. A.K. Gopal has received commercial research support from Spectrum. D.G. Maloney is a consultant/advisory board member of Roche/Genentech. R. Fisher is a consultant/advisory board member of Micromet. No potential conflicts of interest were disclosed by the other authors.

**Authors’ Contributions**

**Conception and design:** O.W. Press, J.M. Unger, J.W. Friedberg, M. LeBlanc, D.G. Maloney, T.P. Miller

**Development of methodology:** O.W. Press, J.M. Unger, M. LeBlanc, T.P. Miller

**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** O.W. Press, L.M. Rimsza, J.W. Friedberg, M.S. Czuczman, M. Kaminski, A.K. Gopal, D.G. Maloney, B.D. Cheson, S.R. Dakhil, T.P. Miller

**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** O.W. Press, J.M. Unger, M. LeBlanc, C. Spier, A.K. Gopal, S.R. Dakhil, T.P. Miller

**Writing, review, and/or revision of the manuscript:** O.W. Press, J.M. Unger, L.M. Rimsza, J.W. Friedberg, M. LeBlanc, M.S. Czuczman, M. Kaminski, R.M. Brezler, A.K. Gopal, D.G. Maloney, B.D. Cheson, S.R. Dakhil, T.P. Miller

**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** O.W. Press, J.M. Unger, L.M. Rimsza

**Study supervision:** O.W. Press, T.P. Miller

**Acknowledgments**

The authors thank Jeri Jardine, Scott Kurruk, Nancy Press, Linda Parker, and the SWOG Operations and Statistical Offices for data management and administrative assistance with the conduct of this trial and the preparation of this article. The authors also acknowledge the generous support of Corixa and GlaxoSmithKline, who provided tositumomab and iodine¹³¹I–tositumomab for this trial, as well as a research grant covering nuclear medicine costs and HAMA testing.

**Grant Support**

This study was supported in part by the following PHS Cooperative Agreement grant numbers awarded by the National Cancer Institute, DHHS: CA21102, CA38926, CA11083, CA27057, CA13612, CA35431, CA20319, CA35090, CA35261, CA46282, CA45560, CA58861, CA5281, CA67575, CA35176, CA45377, CA51287, CA76429, CA22433, CA63844, CA12644, CA55919, CA45450, CA35119, CA3178, CA36416, CA67663, CA4441, CA35192, CA37981, CA76447, CA25654, CA16385, CA073590, CA45808, CA58682, CA14028, CA58723, CA74811, and CA13946 (to SWOG). NCI RO1 CA676287 (to O.W. Press), and in part by a grant from Corixa and GlaxoSmithKline (to SWOG).

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 April 30, 2013; revised August 22, 2013; accepted September 16, 2013; published OnlineFirst October 15, 2013.


A Comparative Analysis of Prognostic Factor Models for Follicular Lymphoma Based on a Phase III Trial of CHOP–Rituximab versus CHOP + $^{131}$Iodine—Tositumomab

Oliver W. Press, Joseph M. Unger, Lisa M. Rimsza, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-13-1120

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2013/10/15/1078-0432.CCR-13-1120.DC1

Cited articles
This article cites 34 articles, 23 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/19/23/6624.full.html#ref-list-1

Citing articles
This article has been cited by 5 HighWire-hosted articles. Access the articles at:
/content/19/23/6624.full.html#related-urls

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.