Risk prediction for patients with follicular lymphoma and chronic lymphocytic leukemia

Vindi Jurinovic

2018

Aus dem Institut für Medizinische Informationsverarbeitung,
Biometrie und Epidemiologie
der Ludwig-Maximilians-Universität München
Direktor: Prof. Dr. rer. nat. Ulrich Mansmann

Risk prediction for patients with follicular lymphoma and chronic lymphocytic leukemia

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät der
Ludwig–Maximilians–Universität zu München

vorgelegt von
Vindi Jurinovic
aus Sinj, Kroatien

2018

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. Ulrich Mansmann
Mitberichterstatter:	Prof. Dr. Michael Ewers Prof. Dr. Irmela Jeremias
Dekan:	Prof. Dr. med. dent. Reinhard Hickel
Tag der mündlichen Prüfung:	11.04.2018

Eidesstattliche Versicherung

Jurinovic Vindi

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Thema

Risk prediction for patients with follicular lymphoma and chronic lymphocytic leukemia

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 06.08.2017

Ort, Datum

Unterschrift Doktorandin

Contents

Nomenclature

CLL	Chronic lymphocytic leukemia
FFS	Failure free survival
FL	Follicular lymphoma
FLIPI	Follicular Lymphoma International Prognostic Index
IgVH	Immunoglobulin heavy chain variable region
Lasso	Least absolute shrinkage and selection operator
LDH	Lactate dehydrogenase
NHL	Non-Hodgkin lymphoma
OLS	Ordinary least squares
OS	Overall survival
PC	Principal component
PCA	Principal component analysis
PCR	Principal component regression
POD24	Treatment outcome at 24 months from treatment initiation
qRT-PCR	Real-time quantitative reverse transcription polymerase chain reaction
SLL	Small lymphocytic lymphoma

List of Publications

This cumulative dissertation is a summary of the following three publications. In the remainder of this thesis, they will be referred to as the first, second, and third publication, respectively. A part of the second publication was presented in a talk at the 57th annual meeting of the American Society of Hematology.

Pastore A*, **Jurinovic V***, Kridel R*, Hoster E* et al. Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. *The Lancet Oncology*, 16(9):1111-1122, 2015.

Jurinovic V, Kridel R, Staiger AM, et al. Clinicogenetic risk models predict early progression of follicular lymphoma after first-line immunochemotherapy. *Blood*, 128(8):1112-1120, 2016.

Herold T, **Jurinovic V**, Metzeler KH, et al. An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. *Leukemia*, 25(10):1639-1645, 2011.

Conference abstract

Jurinovic V, Kridel R, Staiger AM, et al. A Clinicogenetic Risk Model (m7-FLIPI) Prospectively Identifies One-Half of Patients with Early Disease Progression of Follicular Lymphoma after First-Line Immunochemotherapy. *Blood*, 126(23):333-333, 2015.

*Equal contribution

Summary

Introduction
Follicular lymphoma (FL) and chronic lymphocytic leukemia (CLL) have long been considered indolent diseases, characterized by a slow progression and high response rates to initial therapy. We now know that the clinical course of both diseases is highly heterogeneous, with a subset of patients experiencing rapid disease progression or resistance to initial treatment. The most commonly used risk classifiers in both FL and CLL are based solely on clinical variables and show marked variation in clinical course within the risk categories. The aim of this work was to improve risk prediction in FL and CLL by incorporating molecular biomarkers into comprehensive multivariable risk models.

Methods
For FL, we analyzed the mutational status of 74 genes in 151 patients uniformly treated within the GLSG2000 trial of the German Low-Grade Lymphoma Study Group. This cohort was used as a training set to develop risk models for prediction of failure free survival (FFS) and treatment failure within 24 months from treatment initiation (POD24). The models were tested on a validation cohort consisting of 107 patients from a population-based registry of the British Columbia Cancer Agency. All patients had symptomatic, advanced stage or bulky FL considered ineligible for potentially curative irradiation. In addition to model validation, the models were extensively compared to each other and the Follicular Lymphoma International Prognostic Index (FLIPI), the most widely used risk classifier for patients with FL.
For CLL, the training cohort consisted of gene expression profiles of 151 samples measured by whole-genome microarrays. The cohort was used to develop a model for prediction of overall survival (OS). For validation on a independent cohort of 149 patients, the expression of genes from the final model was assayed by real-time quantitative reverse transcription PCR. The Laboratory for Leukemia Diagnostics of the Ludwig Maximilian University of Munich received the samples from german hospitals and private practitioners for routine diagnostics.

Results
We developed and independently validated a new risk classifier for FL, termed m7-FLIPI, that consisted of two clinical variabes and the mutational status of seven genes. The m7-FLIPI was highly predictive in both the training and the validation cohort and outperformed the predictive value of FLIPI alone. Furthermore, we could show that the pre-treatment classifier m7-FLIPI is able to predict POD24, which is a strong predictor for OS, but of limited clinical utility as it can not be assessed prior to treatment. To further increase the power of predicting POD24, we developed a new risk model with POD24 as outcome. The new classifier, POD24-PI, was comprised of four variables which were all contained within the m7-FLIPI. POD24-PI was more sensitive than the m7-FLIPI in predicting POD24, however at a cost of a lower specificity. In comparison with FLIPI and POD24-PI, the m7-FLIPI emerged as the most stringent classifier with the highest accuracy to predict POD24.
For patients with CLL, we developed a model for prediction of OS composed of expression levels of only eight genes. The new risk score, termed PS.8, was highly significant and remained independent of established prognostic markers in a multivariable Cox-model. Furthermore, it also significantly predicted shorter OS and shorter time to treatment in patients with Binet A stage CLL.

Conclusion
We were able to significantly improve risk prediction for patients with FL by designing comprehensive risk models that integrate molecular biomarkers which reflect disease biology. The m7-FLIPI is currently the most promising risk classifier that can identify the smallest group of patients at highest risk of early treatment failure. Though POD24-PI is less specific, its higher sensitivity may make it valuable when testing well-tolerated treatments. Subsequent studies are on the way that will further test the m7-FLIPI and explore its predictive potential also in the context of other regimens. If validated, the m7-FLIPI can be a useful tool to select trial candidates for novel treatments, as standard immunochemotherapy is currently of limited benefit for high-risk patients.
For CLL patients, we could improve the prediction of OS by constructing a model with less than ten genes from a whole-transcriptome data set. The fact that the gene expression in the validation cohort was measured on a different technical platform demonstrates the robustness of the model. Because of the small number of its components, PS.8 can easily be performed in routine diagnostics if validated in other studies.

Zusammenfassung

Einführung
Das follikuläre Lymphom (FL) und die chronische lymphatische Leukämie (CLL) wurden lange Zeit für indolente Krankheiten gehalten, die durch einen langsamen Verlauf und ein gutes Ansprechen auf die initiale Therapie charakterisiert sind. Heute wissen wir, dass beide Krankheiten einen sehr heterogenen Verlauf haben, wobei eine Gruppe der Patienten durch ein rasches Fortschreiten der Krankheit oder primäre Therapieresistenz charakterisiert ist. Die am häufigsten verwendeten Risikoklassifikatoren in FL und CLL sind aus rein klinischen Variablen zusammengesetzt und zeigen innerhalb der Risikogruppen einen merklichen Unterschied im Krankheitsverlauf. Das Ziel dieser Arbeit war die Verbesserung der Risikovorhersage für Patienten mit FL und CLL durch Entwicklung von umfassenden multivariablen Risikomodellen mit Hilfe von molekularbiologischen Daten.

Methoden
Wir untersuchten den Mutationsstatus von 74 Genen in 151 FL Patienten, die innerhalb der GLSG2000 Studie der Deutschen Studiengruppe für niedrigmaligne Lymphome einheitlich behandelt wurden. Diese Patientenkohorte wurde als Trainingskohorte für die Entwicklung der Prädiktionsmodelle für die Zeit bis zum Therapieversagen (FFS) und das Therapieversagen innerhalb von 24 Monaten (POD24) verwendet. Die Modelle wurden auf einer unabhängigen Validierungskohorte von 107 Patienten aus dem populationsbasierten Register der British Columbia Cancer Agency getestet. Alle Patienten hatten ein symptomatisches, fortgeschrittenes FL, oder konnten wegen hoher Tumorlast nicht kurativ bestrahlt werden. Zusätzlich zur Modellvalidierung wurden die Modelle ausgiebig miteinander und dem meistverbreiteten Risikoklassifikator für FL Patienten, dem Follicular Lymphoma International Prognostic Index (FLIPI), verglichen.
Für das CLL Projekt bestand die Trainingskohorte aus Genexpressionsprofilen von 151 Proben, die mit Gesamtgenom-Microarrays analysiert wurden. Auf dieser Kohorte wurde ein prädiktives Modell für das Gesamtüberleben (OS) entwickelt. Für die Validierung auf einer unabhängigen Kohorte von 149 Patienten wurde die Expres-

sion der Gene aus dem finalen Modell mit quantitativer Echtzeit-Reverse-Transkriptase-PCR gemessen. Das Labor für Leukämiediagnostik der Ludwig-Maximilians-Universität München erhielt die Proben von deutschen Krankenhäusern und Privatpraxen für Routinediagnostik.

Ergebnisse
Wir entwickelten und validierten einen neuen Risikoklassifikator für FL, genannt m7-FLIPI, der aus zwei klinischen Variablen und dem Mutationsstatus von sieben Genen zusammengesetzt ist. Der m7-FLIPI war sowohl in der Trainings- als auch in der Validierungskohorte hoch prädiktiv und übetraf den prädiktiven Wert von FLIPI. Darüberhinaus konnten wir zeigen dass der prätherapeutische m7-FLIPI auch für POD24 prädiktiv ist. POD24 ist ein starker Prädiktor für das OS, aber von bedingtem Nutzen, da er nicht vor der Therapie bestimmt werden kann. Um die Power der Prädiktion von POD24 weiter zu erhöhen, entwickelten wir ein neues Modell mit POD24 als Zielvariable. Der neue Risikoklassifikator, POD24-PI, bestand aus vier Variablen, die allesamt im m7-FLIPI enthalten sind. Im Vergleich mit FLIPI und POD24-PI zeigte sich der m7-FLIPI als der spezifischste Klassifikator mit der höchsten Prädiktionsgenauigkeit für POD24.
Für Patienten mit CLL entwickelten wir einen neuen Algorithmus für die Vorhersage des OS, bestehend aus den Expressionswerten von lediglich acht Genen. Der neue Risikoscore, genannt PS.8, war hochsignifikant und blieb im multivariablen Cox-Modell unabhängig von etablierten prognostischen Markern. Darüberhinaus war er in der Untergruppe der Patienten im Binet A Stadium sowohl für das OS als auch für die Zeit bis zum Therapiebeginn signifikant prädiktiv.

Schlussfolgerung
Durch Entwicklung von Risikomodellen mit Hilfe von molekularen Biomarkern, die die Krankheitsbiologie widerspiegeln, konnten wir die Risikovorhersage für Patienten mit FL signifikant verbessern. Der m7-FLIPI ist gegenwärtig der vielversprechendste Risikoklassifikator, der die kleinste Gruppe der Patienten mit höchstem Risiko für frühes Therapieversagen identifizieren kann. Obwohl POD24-PI weniger spezifisch ist, kann er durch seine höhere Sensitivität beim Testen von gut verträglichen Therapien nützlich sein. Für die Validierung des m7-FLIPI sind schon weitere Studien geplant, die sein prädiktives Potential auch im Kontext anderer Therapien untersuchen werden. Wenn der m7-FLIPI validiert wird, kann er ein hilfreiches Werkzeug für die Auswahl der Studienteilnehmer für das Testen neuer Therapien sein, da die gängige Immun-Chemotherapie für ein Hochrisiko-Kollektiv nur von eingschränktem Nutzen ist.
Für CLL Patienten konnten wir die Vorhersage des OS verbessern, indem wir aus einem Gesamt-Transkriptom Datensatz ein Modell mit weniger als zehn Genen kon-

struierten. Die Tatsache, dass die Genexpression in der Validierungskohorte auf einer anderen technischen Plattform gemessen wurde, spricht für die Robustheit des Modells. Durch die kleine Anzahl seiner Komponenten kann PS.8 leicht in die Routinediagnostik übernommen werden, falls er in anderen Studien validiert werden sollte.

1. Introduction

1.1 Follicular Lymphoma

Follicular lymphoma (FL) is one of the most common non-Hodgkin lymphomas (NHL) worldwide with a median age of about 60 years at time of diagnosis [1]. FL is characterized by a highly variable clinical course with some patients living more than a decade without need of treatment, while others develop early progression of disease and have a significantly shorter overall survival (OS) [2]. About 85% of cases harbor the t(14;18)(q32;q21) translocation that is considered a hallmark of FL [3]. The tumor originates from germinal center B cells and resembles follicles consisting of variable proportions of centrocytes and centroblasts, surrounded by a mixture of non-malignant cells [3]. The proportion of centrocytes to centroblasts gives rise to the grading scheme of FL. Staging is done according to the Ann-Arbor staging system for malignant lymphomas that classifies the tumor depending on its location [4].

About 15 – 25% of patients are diagnosed at early stage I or II that may potentially be cured with radiotherapy [5]. Advanced stage disease is currently considered incurable and the treatment goals focus on prolonged failure free (FFS) and overall survival. Since the addition of the monoclonal CD20-antibody rituximab to various first-line chemotherapy regimens has been shown to substantially improve the outcome of FL patients [6–8], the combination of rituximab and chemotherapy has become a common approach for frontline treatment of advanced stage FL. Such immunochemotherapies are able to induce an overall response rate of more than 90% in advanced stages requiring treatment [5]. However, 20 – 25% of patients are primary treatment resistant or experience treatment failure within 24 months of initial im-

munochemotherapy (POD24) [9]. These high-risk patients should be identified prior to treatment since they clearly do not benefit from standard therapy and should be prioritized for alternative, highly active regimens.
The most widely used prognostic index in patients with FL is the Follicular Lymphoma International Prognostic Index (FLIPI) [10]. It is comprised of 5 risk factors:

- age > 60 years
- hemoglobin level < 120g/L
- elevated serum lactate dehydrogenase (LDH) level
- Ann-Arbor stage III/IV
- > 4 involved nodal areas

Based on the number of positive risk factors, the FLIPI divides patients into three risk groups:

Risk group	**Number of risk factors**	**Proportion of patients** (%)	**5-year OS** (%)	**10-year OS** (%)
Low	≤ 1	36 - 47	81 - 91	62 - 71
Intermediate	2	26 - 40	66 - 78	48 - 51
High	≥ 3	27	47 - 53	34 - 36

Table 1.1: Prognostic groups in FL according to FLIPI [10, 11].

Though it was developed to predict overall survival for untreated patients in the pre-rituximab era, the prognostic value of FLIPI has been confirmed for patients treated with immunochemotherapy [12] and patients in first relapse [13]. It has also been shown to predict FFS, with the low and intermediate risk group falling into one category [12]. However, FLIPI classifies about one-half of patients requiring treatment as having a high-risk disease, but only about one third of them eventually experience early treatment failure [5]. Because of this poor specificity, FLIPI is not routinely used to guide treatment decisions. Thus, risk prognostication in FL needs to be further improved to capture the high-risk group of patients with insufficient or short response to treatment.
In addition to clinical risk factors, a number of gene mutations has recently been reported to have an association with treatment outcome [3, 14–16]. However, there has

been no attempt to combine different genetic mutations into one predictive model, as it was done with clinical variables in case of FLIPI. The goal of this work was to improve risk prediction for patients with FL by combining information on recurrent gene mutations and clinical risk factors into one predictive model.
In the first publication, we developed and independently validated a new clinicogenetic risk model for FL, which we termed the m7-FLIPI. In the same year, POD24 has been shown to be a highly significant surrogate point for overall survival [9]. POD24 is a post-treatment variable and thus can not be used to guide risk-adapted frontline treatment decisions. This gave rise to the question if the pre-treatment classifier m7-FLIPI can also predict POD24. This issue was addressed in the second publication. Additionally, we developed and validated a new predictive model specifically designed to predict POD24 and compared its performance to FLIPI and m7-FLIPI.

1.2 Chronic Lymphocytic Leukemia

Chronic lymphocytic leukemia (CLL) is the most common type of leukemia in adults, affecting mostly elderly patients with a median age of 65 years at time of diagnosis [17]. Although termed *leukemia*, the World Health Organization considers CLL and small lymphocytic lymphoma (SLL) as different manifestations of the same disease, the only difference being in the location of malignant cells. CLL cells are located predominantly in the blood and bone marrow, while SLL mostly affects the lymph nodes. Like FL, CLL is thus considered a form of NHL.
A major feature CLL shares with FL is its clinical and biological heterogeneity. While some patients have a normal life span, others experience a rapid progression with an extremely poor outcome [17]. The most widely used risk classification for CLL in Europe is the Binet staging system that categorizes patients into three prognostic groups [18]. Binet stage A and B are both defined by the absence of anemia and thrombocytopenia, whereby the additional presence of 3 or more involved nodal areas defines Binet stage B. Binet stage C is defined by the presence of anemia and/or thrombocytopenia, irrespective of the number of involved areas. In the US and Canada, the Rai staging system is more frequently used [19]. Originally,

it classified patients into 5 categories, but has since been revised to reduce the number of prognostic groups to 3 [20]. Thus, both systems categorize patients into a low-, intermediate- and a high-risk group, albeit a substantial proportion of patients is classified differently by the two systems. The size of different risk categories as defined by the Binet and Rai systems and the corresponding median OS are shown in table 1.2.

Risk group	Binet			Rai		
	Stage	Proportion (%)	Median OS (years)	Stage	Proportion (%)	Median OS (years)
Low	A	63	> 10	0	30	> 10
Intermediate	B	30	5	1 2	60	7
High	C	7	1.5	3 4	10	1.5

Table 1.2: The Binet and Rai staging systems [21].

The categories defined by the two systems do not overlap well with one another, with Binet stage A being about twice as large as Rai low-risk group. Furthermore, there is substantial heterogeneity in the disease course within the risk categories. About 40% of patients with Binet stage A experience disease progression to advanced stages, and about 25% die of causes related to CLL [22]. Thus, new prognostic markers are needed that can capture the heterogeneity in outcome within the different stage categories.

Among the first biomarkers shown to be predictive of OS were the somatic mutations in the immunoglobulin heavy chain variable region (IgVH) genes [23, 24]. Patients with CLL cells that use unmutated IgVH genes have inferior survival compared to patients with mutated IgVH genes. Additionally, different chromosomal aberrations have been associated with the disease course [25–28]. Deletions involving the long arm of chromosome 13 (del(13q)) represent the most common chromosomal aberration in CLL, occuring in about 55% of cases. A sole del(13q) is characterized by a benign disease course with a median survival of more than 10 years. The second most common aberration, del(11q), is more common among younger patients and is associated with inferior outcome. Other frequent aberrations in CLL include trisomy 12q, del(6q) and del(17p), which is associated with resistance to treatment and poorest survival with a median OS of less than three years. Chromosomal aberra-

tions have by now been incorporated into a hierarchical model that can significantly predict overall survival in CLL patients [29].
While the predictive value of chromosomal aberrations has been well studied, gene expression levels have not been systematically analyzed in connection with OS. The objective of our work that resulted in the third publication was to improve the prognosis of overall survival by using the information on the whole transcriptome of CLL cells and incorporating the gene expression levels into one single model. While we used oligonucleotide microarrays to develop the model, the gene expression in the validation cohort was measured with real-time quantitative reverse transcription PCR (qRT-PCR), making the validation even more stringent. Importantly, the model was also validated in the subgroups defined by other relevant biomarkers, including the largest Binet risk group of patients classified as having a low-risk disease.

1.3 Statistical Methods

1.3.1 Preliminaries

The terms *prognostic* and *predictive* marker are often used interchangeably to describe an association between a predictor and the outcome. However, there is a substantial difference between a prognostic and a predictive variable. A prognostic marker is associated with the outcome irrespective of the therapy a patient receives. In contrast, a predictive marker is associated with the outcome only in the context of a certain treatment. On the other hand, in statistical terminology we say that a marker M is predictive of the outcome Y if there is an association between M and Y. However, this does not mean that M is a predictive marker in a medical sense. To explore this, we would need to test the association between M and Y for various treatment regimens, for example by including an interaction term between the marker and the treatment variable in a regression model. Thus, there is a certain difference between the statistical term M *is predictive of* Y, and the term *predictive marker* that is often used in medical literature.
In the first two publications that deal with risk prediction in FL, we developed two

risk models for prediction of FFS and POD24. In the third publication, a model for prediction of overall survival in patients with CLL was developed. In the first two projects, the set of predictor variables was less than 100, while the number of potential predictors in the third project exceeded 50 thousand. Moreover, while the m7-FLIPI and PS.8 were developed to predict a survival variable, POD24-PI was designed for the binary outcome POD24. Different types of outcomes and different number of predictors mandate different approaches in developing predictive models. In the following, the methods used for model building in the three cases will be introduced.

Linear Regression

Let $\mathbf{X} = (\mathbf{x}^0, \ldots, \mathbf{x}^p)$ be a matrix of $p+1$ predictor variables with sample size n, $\mathbf{y} = (y_1, \ldots, y_n)^T$ the outcome (response) of interest. The best known and most frequently used approach for predicting a normally distributed outcome is the linear regression model, where the association between the response and the predictors is assumed to be linear, $\mathbf{y} = \beta_0\mathbf{x}^0 + \ldots + \beta_p\mathbf{x}^p = \mathbf{X}\boldsymbol{\beta}$. From this equation, we want to calculate the regression coefficients $\beta_0, \ldots, \beta_p$ that can later be used to predict outcome for future subjects. In the usual setting, however, this system is overdetermined ($n > p+1$) and has no solution. Thus, the coefficients are calculated so that the predicted values for the data set $\mathbf{X}$ are as close as possible to the true values $\mathbf{y}$. This is done by minimizing the residual sum of squares,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_j^i\beta_j\right)^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

If the vectors $\mathbf{x}^0, \ldots, \mathbf{x}^p$ are linearly independent, the problem has a unique solution $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p) = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ called the *ordinary least squares (OLS) estimates*. Usually, $\mathbf{x}^0$ is set to a constant value $x_i^0 = 1$ for all $i \in \{1, \ldots, n\}$ and the corresponding β_0 is called the *intercept*.
A simple linear model for a simulated data set with sample size 500, $X \sim \mathcal{N}(0,1)$ and $Y \sim X + \mathcal{N}(0,1)$ is shown in figure 1.1. The estimated model is $\hat{Y} = -0.06 + 1.08 \cdot X$, showing the power of linear regression in case of normally distributed outcomes.

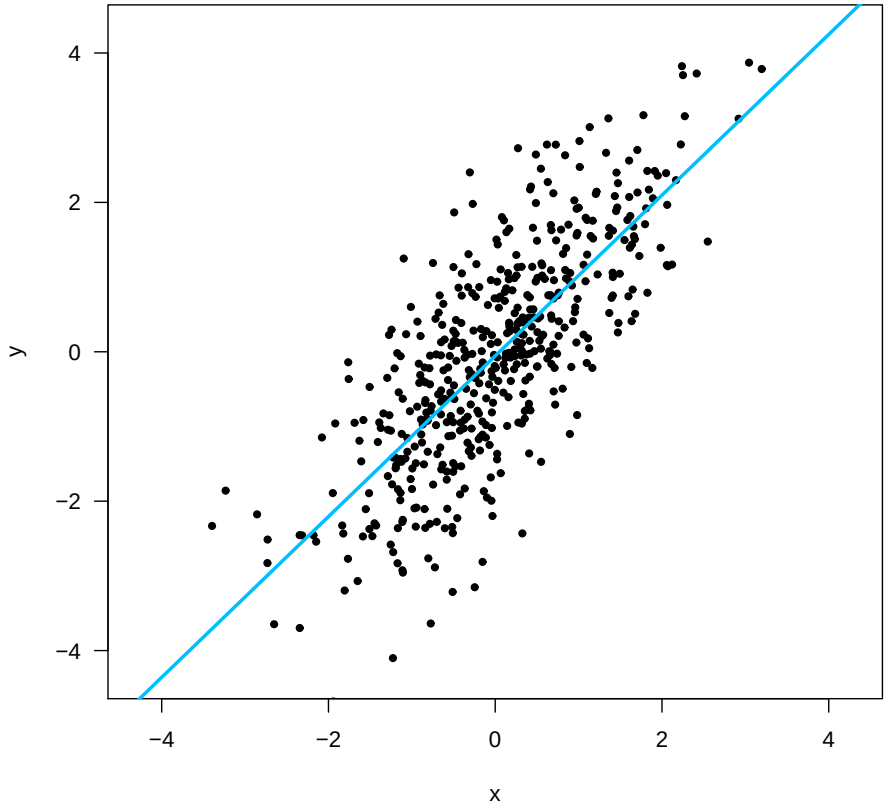

Figure 1.1: A simple linear model with one predictor variable.

Logistic Regression

As can be seen from figure 1.1, the linear model allows for any real value of Y to be predicted. This is reasonable for a normally distributed outcome, but in some cases, certain values of Y are not possible. For example, a count variable is always non-negative, while a binary variable can only take on two values. When the linear model is applied to that kind of data, it predicts values that can not occur in reality. Figure 1.2 shows the association for a simulated data set between a binomially distributed binary outcome $Y \sim \mathrm{B}(500, 0.5)$ and a predictor $X \sim \mathcal{N}(0, 1)$ for $Y = 0$ and $X \sim \mathcal{N}(2, 1)$ in case of $Y = 1$. In figure A, the blue line shows the values for Y as predicted by the linear model. For example, the predicted value for

$X = -2$ is $\hat{Y} = -0.27$, which is difficult to interpret since this value is not permissible. Therefore, linear regression is not suitable for this kind of data and a model that predicts reasonable and interpretable values is needed. The logistic regression model is the most widely used approach for prediction of a binary outcome. It associates the linear combination of predictors not with the outcome, but with the logarithm of its odds:

$$\log(\text{Odds}) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \mathbf{x}^1 + \ldots + \beta_p \mathbf{x}^p = \mathbf{X}\boldsymbol{\beta},$$

where π is the probability that the event of interest will occur. Solving for π yields

$$\pi = \frac{1}{1 + \exp(-\mathbf{X}\boldsymbol{\beta})}.$$

The positiveness of the exponential function ensures that the estimate for π will always be in the interval $]0, 1[$, making the results of logistic regression meaningful and easily interpretable. Figure 1.2 B shows the values for π as predicted by the logistic regression model for the previous example. This time, the predicted value for $X = -2$ is $\pi = 0.0019$, and there are no predicted values outside the interval $[0, 1]$.

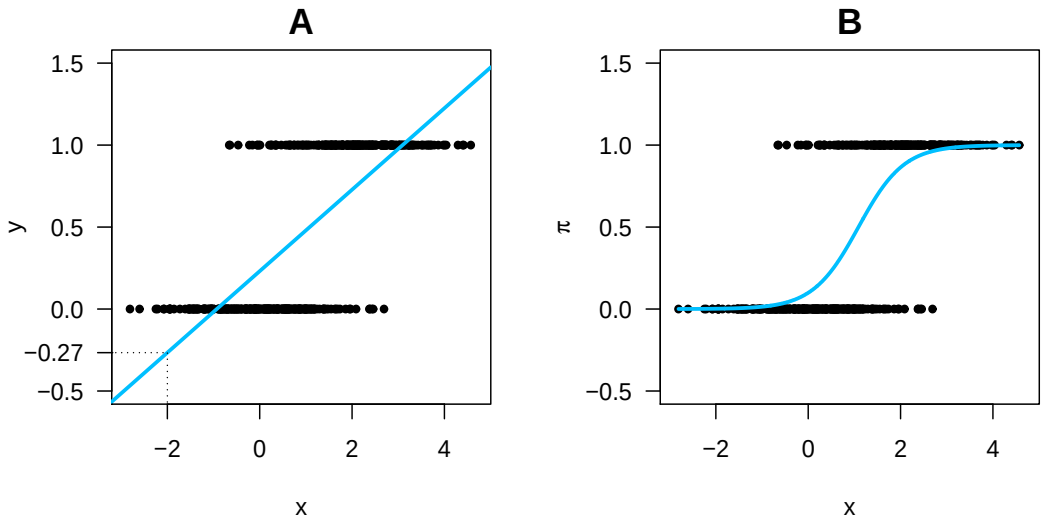

Figure 1.2: Association between a predictor and a binary outcome, and the predicted values by A: the linear model; B: the logistic model.

The regression coefficients are chosen so that they maximize the joint probability (likelihood) for the observed outcomes of all n subjects. The estimation is not as straightforward as in the case of linear regression, as it involves solving $p + 1$ nonlinear equations that can not be done algebraically. Thus, the solutions need to be estimated iteratively, which in practice is usually done with the Newton-Raphson method for solving nonlinear equations [30].

The Cox Proportional Hazards Model

Survival data have the form $(\mathbf{x}_j, y_j, \delta_j)$, where $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})$ is a vector of p predictor values for the j–th patient, y_j the observed survival time and δ_j the indicator for the event (1 = event, 0 = censored). The most popular regression model for survival data is the Cox proportional hazards model [31], where the hazard $h(t)$ at time t for individual j is modeled as

$$h(t|\mathbf{x}_j) = h_0(t) \cdot \exp(\beta_1 x_{j1} + \ldots + \beta_p x_{jp}),$$

where $h_0(t)$ is the baseline hazard for individuals with all predictor values equal to zero. It is usually unspecified and the regression coefficients can be estimated without any information on $h_0(t)$. Similar to logistic regression, the estimation is done iteratively with the goal of maximizing the joint probability of all observed events.

1.3.2 Lasso

In general, all regression coefficients estimated with previously introduced models will be non-zero, even if the corresponding predictor variable is not associated with the outcome. This makes the interpretability of results increasingly difficult with growing number of predictor variables. Additionally, while their bias is usually low, the estimates often have a large variance. To overcome these two problems, a new linear regression method called *Lasso* (least absolute shrinkage and selection operator) was introduced by Robert Tibshirani in 1996 [32]. Assuming the outcome

is normalized to have a mean value equal to zero and all predictors are standardized so that $\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j^i = 0$ and $\frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{x}_j^i\right)^2 = 1$, Lasso minimizes

$$\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \text{ subject to } \sum_{i=1}^{p}|\beta_i| \leq t.$$

Thus, the regression coefficients are penalized by the constraint that the sum of their absolute values should not exceed the value t. Such regression models that set constraints on coefficient values are called *penalized regression models*.

In case of Lasso, a value of t smaller than the sum of absolute values of OLS estimates will cause the regression coefficients to be shrunk towards zero, with some values being exactly zero if t is sufficiently small. Therefore, Lasso increases the interpretability of data by allowing only the most relevant predictors to enter the model. In practice, the amount of shrinkage needed for a specific case is usually estimated with cross-validation. Penalization of regression coefficients has since been extended to other regression models, including logistic [33] and Cox proportional hazards regression [34].

We used penalized regression for development of both the m7-FLIPI and the model for prediction of POD24. As m7-FLIPI was designed to predict failure free survival, it was built with Lasso for the Cox proportional hazards model. In case of a binary outcome POD24, penalized logistic regression was used for the construction of POD24-PI.

The Lasso path plot in figure 1.3 shows the association between the amount of shrinkage controlled by the parameter λ and the Lasso regression coefficients for the penalized variables used to build the m7-FLIPI. On the far left, the amount of shrinkage is high and all Lasso coefficients are equal to 0. With λ decreasing, the first variable to enter the model is EZH2, followed by ARID1A and EP300. The dashed line indicates the optimal λ estimated by 10-fold cross-validation that resulted in the m7-FLIPI. Thus, all variables that entered the model on the right side of $\lambda_{optimal}$ did not make it into the m7-FLIPI.

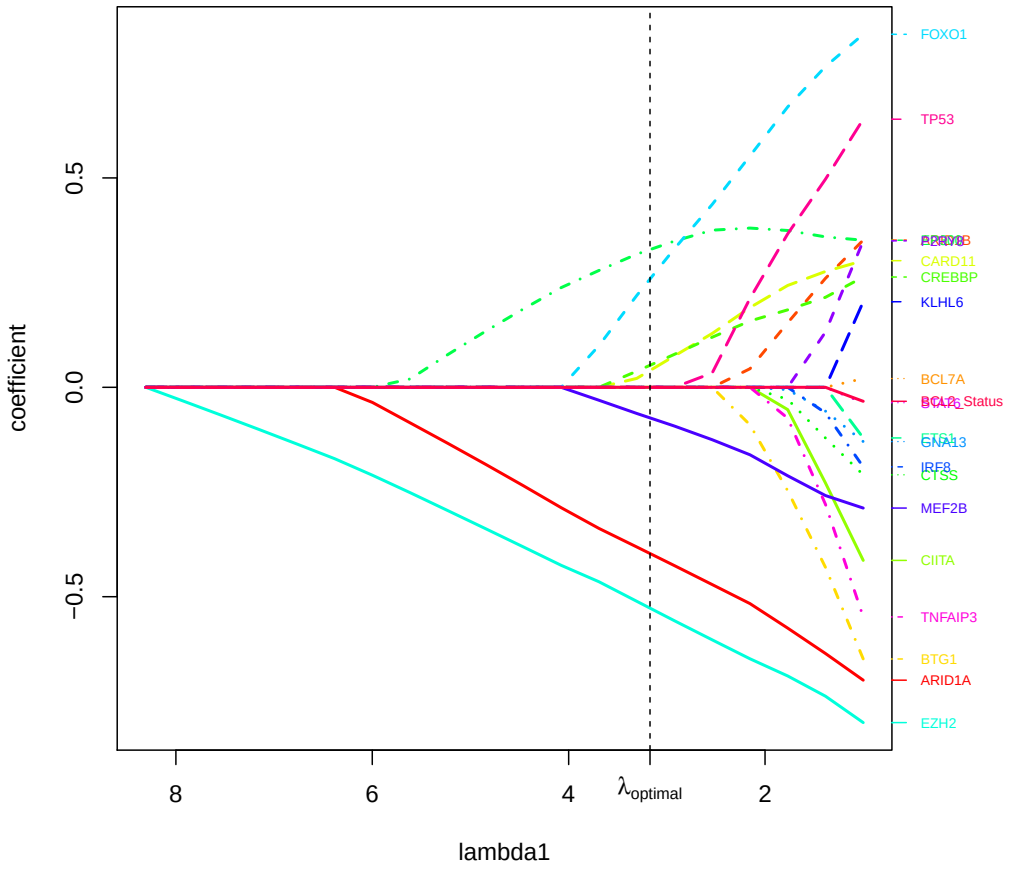

Figure 1.3: Lasso path plot for the FL training cohort with FFS as outcome.

1.3.3 Bootstrap

The bootstrap is a resampling procedure that was introduced as a method to estimate the sampling distribution of a data set [35]. In statistics, we usually use a subsample from the population of interest to answer questions about the whole population. Suppose we want to estimate the mean value of variable X in the population by randomly sampling 100 individuals. Each time we repeat this experiment, we will get a different value for the mean $\bar{X}$ because of the random nature of sampling. If we could repeat the procedure infinitely often, we would get all possible

estimates for $\bar{X}$. The distribution of these values is called the *sampling distribution*. The idea behind bootstrapping is to simulate the many random samples we would (but usually can not) draw from the whole population by sampling many times from one single subsample. A bootstrap sample is created by drawing n observations with replacement from the original data set of sample size n. This means that a bootstrap sample will contain some observations multiple times, while others will not be selected at all. This procedure is repeated B times and the estimate of interest is calculated for each of the B bootstrap samples. The distribution of these estimates provides an approximation for the sampling distribution, whose variance gives us the information on the reliability of the estimate from the original data set.

Bootstrapping can also be used in combination with regression models to improve the stability of results. To test if a predictor is associated with the outcome, we can fit a regression model in each bootstrap sample instead of testing the association once in the original data set. A robust predictor will in general not depend much on random changes in data and will remain significant in most bootstrap samples. The results from all regression models can be used to interpret each predictor's significance. For example, we can say that a predictor is associated with the outcome only if it is significant in more than 50% of bootstrap samples, or if the mean of bootstrap regression coefficients is larger than some specified value. This way we can ensure that only robust predictors enter our further analyses.

Another way to use bootstrap is to estimate the amount of overfitting that is a common problem in regression models [36]. Overfitting describes the fact that in general, a score designed to predict the outcome Y will have a higher predictive power in the data set it was developed on (training set) than in any other data set (validation set). This means that in a regression model with dependent variable Y, the score will have a higher regression coefficient in the training than in the validation set. The difference between these two coefficients is called the *optimism*, which can be used as a measure for overfitting. To estimate the optimism for score S, the same procedure used to construct S is repeated on each bootstrap sample of the training set, resulting in B different scores. Next, the training set is used as a validation set for each of the B scores obtained by bootstrapping. The average across the B optimism values is the estimated optimism for score S. To correct for overfitting, the optimism is subtracted from the overestimated regression coefficient

the score S has in the training set.
We used bootstrapping for different purposes in all three publications. For prediction of survival in CLL patients where we had to build a model from more than 50 000 variables, we applied bootstrapping for preselection of the most robust predictors. In the two publications on FL, bootstrapping was used as a method to compare competing models by correcting their raw coefficients for overfitting.
The benefit of bootstrapping combined with Lasso was illustrated in detail in a grant application we submitted to the Wilhelm Sander-Stiftung in 2013. The application under the name *Entwicklung und Validierung einer auf Genexpression basierten Klassifikation zur Prädiktion des Ansprechens der Behandlung für Patienten mit akuter myeloischer Leukämie* was approved in 2013 and the project is currently being finalized. The section from the application considering bootstrapping and Lasso is included in the Appendix.

1.3.4 Supervised Prinicipal Components

Principal component analysis (PCA) was first introduced in 1901 [37] and has since been further developed and extensively used in various scientific fields, from image analysis to electrical engineering. The goal of PCA is to simplify a multivariate data set by uncovering its underlying structure and using this information to reduce its dimension. In figure 1.4, a simple data set containing two correlated variables, X and Y, is plotted in the cartesian coordinate system, denoted by the coordinates x and y. To locate a data point in the picture, we need the information on both cartesian coordinates, since the data have substantial variance in both the x- and the y-direction. However, if we used a new coordinate system denoted by the blue coordinates PC1 and PC2, we would be able to find the approximate position of a point by only providing the value for PC1, since there is only a slight variation in direction of PC2. PCA is about finding this new coordinate system where the coordinates, called the *principal components* (PCs), are positioned so that they explain as much variability in the data as possible. The first principal component, PC1, is chosen as the direction where the data set has its highest variance. Each succeeding principal component is chosen to have the largest remaining variance, under the constraint

that all the principal components have to be orthogonal. Mathematically, this is done by the eigenvalue decomposition of the covariance matrix of the data set, $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}.$$

The columns od $\mathbf{Q}$ contain the principal components, while the diagonal matrix $\mathbf{\Lambda}$ contains the variances along the corresponding PCs.
Once the PCs are found, we can calculate what proportion of variability in the data is explained by each PC and use this information to decide how many PCs to keep when analyzing the data. In the previous example, as much as 96% of total variance in the data is explained by PC1.

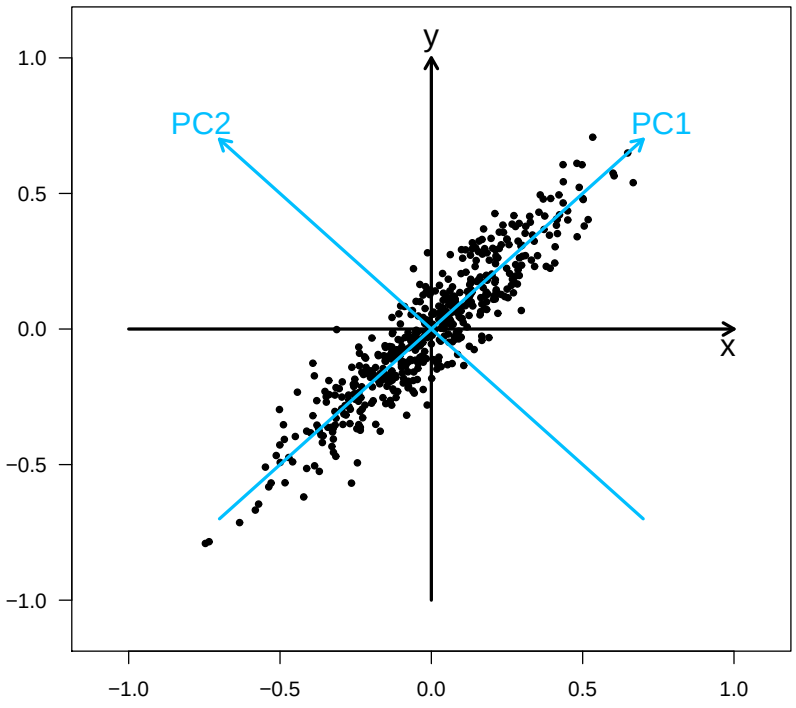

Figure 1.4: Principal components of a two-dimensional data set.

Principal components can be used in previously described regression models, in a technique called *principal component regression* (PCR)[38]. In PCR, principal components of the data set containing the predictors instead of the predictors themselves are used as covariables in the regression model. Usually, a small number of PCs is chosen, depending on the explained variance needed in the particular case. This has a number of advantages. Problems that arise when correlated predictors are present in the model are completely avoided, since the PCs are orthogonal and therefore independent. Furthermore, as only a small number of PCs is usually chosen for regression, overfitting that is caused by using too many predictors can be reduced. However, PCR is not always appropriate and can in certain cases produce misleading results. Suppose the first two PCs have been chosen for regression since they explain the most variability in the data. If these two PCs are not associated with outcome, but PC3 and PC4, the model will not show any association between the data and the outcome. This problem is addressed by supervised principal component analysis [39]. Here, only a subset of the whole data set is chosen for PCA, based on univariate association of each variable with the outcome. The procedure can be simplified as follows:

1. Compute standardized univariate regression coefficients β_i for each variable $\mathbf{x}^i$ in the data set $\mathbf{X}$;
2. Build a subset of $\mathbf{X}$ by taking only those variables with $|\beta_i| > \theta$, where θ is estimated by cross-validation;
3. Do PCA;
4. Do PCR.

This approach will eliminate the possibility of including only irrelevant PCs in the model, since PCA is only done with variables that are associated with the outcome. We chose supervised principal component analysis combined with bootstrapping as a method to construct a predictive model for overall survival in patients with CLL. Step 1. of the algorithm was repeated on 5000 bootstrap samples, and only predictors that were selected in more than 85% of bootstrap samples were chosen for the principal component analysis. Finally, the model was constructed with the first PC as a predictor in a Cox proportional hazards model.

2. Contribution to the Individual Publications

This dissertation summarizes three publications dealing with prediction models in follicular lymphoma and chronic lymphocytic leukemia. In the first publication, we developed and independently validated a new clinicogenetic risk model for FL, the m7-FLIPI. In the second publication, the power of m7-FLIPI to predict POD24 was analyzed, and a new model for prediction of POD24 was constructed. In the third publication, a model for prediction of overall survival in CLL was developed and independently validated.
In the following, each publication and my own contribution to the publication will be outlined.

Publication 1 - Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. *The Lancet Oncology*, 2015.

The aim of this work was to combine clinical variables with the information on the mutational status of 74 recurrently mutated genes to develop a clinicogenetic model for prediction of failure free survival in follicular lymphoma. The model was developed on a training cohort of 151 FL patients treated within the GLSG2000

trial of the German Low-Grade Lymphoma Study Group [8], and validated on an independend data set of 107 patients from a population-based registry of the British Columbia Cancer Agency. The new risk model, termed m7-FLIPI, was highly significant in the validation set and outperformed FLIPI, which so far has been the most widely used prognostic tool for FL.
Supervised by Dr. Eva Hoster and Dr. Michael Unterhalt, I was responsible for the statistical part of the study. This involved the description and statistical analyses of the data sets, the development of the m7-FLIPI, its validation and interpretation. I programmed and executed the R-code for all analyses concerning the description of data, development and validation of the m7-FLIPI and created a number of figures for the final manuscript.

Publication 2 - Clinicogenetic risk models predict early progression of follicular lymphoma after first-line immunochemotherapy. *Blood*, 2016.

This work was a direct result of questions that arose after the first article was published. Shortly after our first publication, an article on POD24 and its strong association with OS was published in the Journal of Clinical Oncology [9]. This prompted us to investigate if m7-FLIPI was also predictive of POD24. The results of this analysis were presented as a talk at the 57th annual meeting of the American Society of Hematology. A suggestion from the audience motivated further work on a development of a new predictive model, specifically designed to predict POD24. This work resulted in a new risk classifier, POD24-PI, that consisted of four risk factors that were all contained within the m7-FLIPI. Finally, FLIPI, m7-FLIPI and the new model POD24-PI were compared to each other in terms of specificity, sensitivity and accuracy, and the pros and cons of each classifier were discussed.
I was responsible for all statistical analyses concerning this publication. This included progamming and executing the whole R code for the data analysis, development and validation of the new model, as well as the analysis and comparison of FLIPI, m7-FLIPI and POD24-PI. Finally, together with Dr. Oliver Weigert, I was involved in creating the figures and writing of the manuscript.

Publication 3 - An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. *Leukemia*, 2011.

In this work, we developed a model for prediction of overall survival in patients with CLL using the gene expression profiles of 151 CLL samples as measured by oligonucleotide microarrays. Although we had more than 50 000 transcript probe sets as potential predictors, we were able to generate a model consisting of the expression values of only 8 genes (PS.8). This model was validated on an independent data set of 149 patients with gene expression measured by qRT-PCR. Furthermore, we showed that the model is predictive of OS and time to treatment in a subgroup of patients that are classified as having a low-risk disease by the widely used Binet staging system.

In this study, I was responsible for all statistical analyses, as well as the programming and executing of the corresponding R code. This included the generation and analysis of bootstrap samples, development and validation of PS.8, and generation of figures for the manuscript. I also performed parallel programming for the analysis of bootstrap samples, which was necessary because of the huge amount of data generated by bootstrapping.

3. Original Publications

3.1 Publication 1 - *The Lancet Oncology*, 2015

Pastore A*, **Jurinovic V***, Kridel R*, Hoster E* et al. Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. *The Lancet Oncology*, 16(9):1111-1122, 2015.

*Equal contribution

Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry

Alessandro Pastore, Vindi Jurinovic*, Robert Kridel*, Eva Hoster*, Annette M Staiger, Monika Szczepanowski, Christiane Pott, Nadja Kopp, Mark Murakami, Heike Horn, Ellen Leich, Alden A Moccia, Anja Mottok, Ashwini Sunkavalli, Paul Van Hummelen, Matthew Ducar, Daisuke Ennishi, Hennady P Shulha, Christoffer Hother, Joseph M Connors, Laurie H Sehn, Martin Dreyling, Donna Neuberg, Peter Möller, Alfred C Feller, Martin L Hansmann, Harald Stein, Andreas Rosenwald, German Ott, Wolfram Klapper, Michael Unterhalt, Wolfgang Hiddemann, Randy D Gascoyne*, David M Weinstock*, Oliver Weigert**

Summary

Background Follicular lymphoma is a clinically and genetically heterogeneous disease, but the prognostic value of somatic mutations has not been systematically assessed. We aimed to improve risk stratification of patients receiving first-line immunochemotherapy by integrating gene mutations into a prognostic model.

Methods We did DNA deep sequencing to retrospectively analyse the mutation status of 74 genes in 151 follicular lymphoma biopsy specimens that were obtained from patients within 1 year before beginning immunochemotherapy consisting of rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP). These patients were recruited between May 4, 2000, and Oct 20, 2010, as part of a phase 3 trial (GLSG2000). Eligible patients had symptomatic, advanced stage follicular lymphoma and were previously untreated. The primary endpoints were failure-free survival (defined as less than a partial remission at the end of induction, relapse, progression, or death) and overall survival calculated from date of treatment initiation. Median follow-up was 7·7 years (IQR 5·5–9·3). Mutations and clinical factors were incorporated into a risk model for failure-free survival using multivariable L1-penalised Cox regression. We validated the risk model in an independent population-based cohort of 107 patients with symptomatic follicular lymphoma considered ineligible for curative irradiation. Pretreatment biopsies were taken between Feb 24, 2004, and Nov 24, 2009, within 1 year before beginning first-line immunochemotherapy consisting of rituximab, cyclophosphamide, vincristine, and prednisone (R-CVP). Median follow-up was 6·7 years (IQR 5·7–7·6).

Findings We established a clinicogenetic risk model (termed m7-FLIPI) that included the mutation status of seven genes (*EZH2, ARID1A, MEF2B, EP300, FOXO1, CREBBP*, and *CARD11*), the Follicular Lymphoma International Prognostic Index (FLIPI), and Eastern Cooperative Oncology Group (ECOG) performance status. In the training cohort, m7-FLIPI defined a high-risk group (28%, 43/151) with 5-year failure-free survival of 38·29% (95% CI 25·31–57·95) versus 77·21% (95% CI 69·21–86·14) for the low-risk group (hazard ratio [HR] 4·14, 95% CI 2·47–6·93; p<0·0001; bootstrap-corrected HR 2·02), and outperformed a prognostic model of only gene mutations (HR 3·76, 95% CI 2·10–6·74; p<0·0001; bootstrap-corrected HR 1·57). The positive predictive value and negative predictive value for 5-year failure-free survival were 64% and 78%, respectively, with a C-index of 0·80 (95% CI 0·71–0·89). In the validation cohort, m7-FLIPI again defined a high-risk group (22%, 24/107) with 5-year failure-free survival of 25·00% (95% CI 12·50–49·99) versus 68·24% (58·84–79·15) in the low-risk group (HR 3·58, 95% CI 2·00–6·42; p<0.0001). The positive predictive value for 5-year failure-free survival was 72% and 68% for negative predictive value, with a C-index of 0·79 (95% CI 0·69–0·89). In the validation cohort, risk stratification by m7-FLIPI outperformed FLIPI alone (HR 2·18, 95% CI 1·21–3·92), and FLIPI combined with ECOG performance status (HR 2·03, 95% CI 1·12–3·67).

Interpretation Integration of the mutational status of seven genes with clinical risk factors improves prognostication for patients with follicular lymphoma receiving first-line immunochemotherapy and is a promising approach to identify the subset at highest risk of treatment failure.

Funding Deutsche Krebshilfe, Terry Fox Research Institute.

Introduction

Follicular lymphoma is the second most common nodal lymphoma worldwide and presents with advanced stage disease in most patients.[1] Several randomised trials have shown that the addition of the monoclonal anti-CD20 antibody rituximab to various first-line chemotherapy regimens improves progression-free survival and overall survival.[2] Additionally, many patients now receive rituximab maintenance after first-line treatment,[2] based on a randomised trial that showed improved progression-free survival.[3]

Lancet Oncol 2015; 16: 1111–22

Published Online August 7, 2015 http://dx.doi.org/10.1016/S1470-2045(15)00169-2

See Comment page 1011

*Contributed equally

Department of Internal Medicine III, University Hospital of the Ludwig-Maximilians-University Munich, Munich, Germany (A Pastore MD, V Jurinovic Dipl Math, E Hoster PhD, Prof M Dreyling MD, M Unterhalt PhD, Prof W Hiddemann MD, O Weigert MD); **Institute for Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-University Munich, Munich, Germany** (V Jurinovic, E Hoster); **Department of Lymphoid Cancer Research and the Centre for Lymphoid Cancer, British Columbia Cancer Agency, Vancouver, BC, Canada** (R Kridel MD, A A Moccia MD, A Mottok MD, D Ennishi PhD, H P Shulha PhD, C Hother MD, Prof J M Connors MD, L H Sehn MD, Prof R D Gascoyne MD); **Department of Clinical Pathology, Robert-Bosch-Krankenhaus, Stuttgart, Germany** (A M Staiger MSc, H Horn PhD, Prof G Ott MD); **Dr Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany** (A M Staiger, H Horn); **University of Tübingen, Tübingen, Germany** (A M Staiger, H Horn); **Haematopathology Section** (M Szczepanowski PhD, Prof W Klapper MD) **and Second Medical Department, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany** (C Pott MD); **Center for Cancer**

Genome Discovery (A Sunkavalli MSc, P Van Hummelen PhD, M Ducar MSc), **Dana-Farber Cancer Institute, Boston, MA, USA** (N Kopp MSc, M Murakami MD, A Sunkavalli, P Van Hummelen, M Ducar, D Neuberg ScD, D M Weinstock MD); **Institute of Pathology, University of Würzburg, and Comprehensive Cancer Center Mainfranken, Würzburg, Germany** (E Leich PhD, Prof A Rosenwald MD); **Institute of Pathology, University of Ulm, Ulm, Germany** (Prof P Möller MD); **Department of Pathology, University Hospital of Schleswig-Holstein, Lübeck, Germany** (Prof A C Feller MD); **Senckenberg Institute of Pathology, Goethe University, Frankfurt am Main, Germany** (Prof M L Hansmann MD); **Berlin Reference Center for Lymphoma and Haematopathology, Berlin, Germany** (Prof H Stein MD); **Clinical Cooperative Group Leukemia, Helmholtz-Center Munich, Munich, Germany** (Prof W Hiddemann, O Weigert); **German Cancer Consortium (DKTK), Heidelberg, Germany** (Prof W Hiddemann, O Weigert); **and German Cancer Research Center (DKFZ), Heidelberg, Germany** (Prof W Hiddemann, O Weigert)

Correspondence to: Dr Oliver Weigert, Department of Internal Medicine III, University Hospital of the Ludwig-Maximilians-University Munich, Marchioninistrasse 15, 81377 Munich, Germany **oliver.weigert@med.uni-muenchen.de**

Research in context

Evidence before this study

On April 9, 2015, we searched the Cochrane Library for all publications without language or date restrictions using the terms "follicular lymphoma" in the title, abstract, keywords, and "prognosis OR prognostic OR prediction OR predictive" anywhere in the text. This search identified a total of 79 publications. Additionally, we searched PubMed for all original publications (ie, review articles were excluded) using the keywords "follicular lymphoma" in the title, and "gene mutation" and "prognostic OR prognosis OR predictive OR prediction" anywhere in the text. This search identified a total of 49 publications.

The relevant papers addressing the effect of molecular alterations on treatment outcome in follicular lymphoma included conflicting reports on the significance of karyotypic abnormalities, *BCL2* translocation breakpoints and immunohistochemistry, positive reports on immunohistochemistry for MUM-1, copy number alterations, uniparenteral loss of heterozygosity, and gene expression signatures, and a negative report on germ-line polymorphisms in single genes (*TP53*, *MDM2*). Additionally, we identified publications on the clinical significance of somatic alterations in single genes (*TP53*, *TNFRSF14*, *CDKN2A*, *BCL2*). In a heterogeneously treated patient population, *TP53* mutations, although rare at initial diagnosis (6%) were predictive for shorter progression-free survival and overall survival when adjusted for International Prognostic Index (IPI). In one study, mutations and chromosomal deletions affecting *TNFRSF14* at 1p36 were associated with high-risk clinical features (ECOG performance status >1, number of extranodal sites >1, high IPI) and shorter overall survival in non-uniformly treated patients, some of whom also received rituximab. In another study, *TNFRSF14* alterations were not associated with poor prognosis. Inactivation of the *CDKN2A* tumour-suppressor gene by deletion or methylation was associated with poor clinical outcome in one study. Coding sequence mutation in *BCL2*, including silent and non-silent variants have recently been associated with increased risk of transformation and shortened overall survival, assumed to represent a surrogate marker for activation-induced cytidine deaminase-mediated genetic instability. Almost all patients in this study were treated in the pre-rituximab era.

Added value of this study

By contrast with previous studies that focused on single gene alterations, we did a multivariable analysis that included a comprehensive compilation of recurrent gene mutations and clinical risk factors in patients with symptomatic follicular lymphoma receiving first-line immunochemotherapy. To the best of our knowledge, this is the largest study to date of relevant mutations in follicular lymphoma, and the first multivariable assessment of their prognostic relevance. The study uses mature clinical data from two independent cohorts of patients: a uniformly treated clinical trial population as a training cohort and a population-based registry as a validation cohort. The resulting model (m7-FLIPI) is, to the best of our knowledge, the first prognostic model in follicular lymphoma that accounts for both clinical factors and genetic alterations. Our study also contrasts with previous studies with regards to the stringent inclusion criteria: all patients had follicular lymphoma grade 1, 2, or 3A confirmed by reference pathology, advanced stage or bulky disease considered ineligible for curative irradiation, symptomatic disease requiring systemic treatment, and a diagnostic biopsy specimen obtained 12 months or less before therapy initiation. Additionally, all patients received a combination of rituximab and chemotherapy (either CVP or CHOP) as first-line treatment. This stringent approach led to the unexpected finding that *EZH2* mutations (found in about one-quarter of patients and associated with a unique transcriptional signature) are linked to favourable outcome in patients with high-risk FLIPI.

Implications of all the available evidence

If validated in subsequent studies, the m7-FLIPI could be highly significant for the medical community, as high-risk patients are clearly underserved by current standard treatment and should be prioritised for innovative treatment options.

Follicular lymphoma is a clinically heterogeneous disease.[4] Identifying the subset of patients at highest risk of early treatment failure is essential, as they are clearly underserved with current standard immunochemotherapy. Prognostication for patients with follicular lymphoma relies primarily on clinical and basic laboratory findings.[4] The Follicular Lymphoma International Prognostic Index (FLIPI) is the most widely used risk model and includes five adverse prognostic factors: age older than 60 years, stage III–IV, haemoglobin less than 120 g/L, four or more involved nodal areas, and elevated serum lactate dehydrogenase.[5] For patients who receive rituximab-containing regimens, the FLIPI can distinguish patients into two risk groups—low or intermediate-risk, and high-risk, for failure-free survival.[6] Whereas about half of patients requiring systemic treatment are classified as high-risk FLIPI, only a third of these will experience treatment failure within 2 years after treatment initiation.[6] Although useful, the FLIPI needs to be further improved to identify those patients at highest risk of early treatment failure. Currently, the FLIPI is not routinely used to guide risk-adapted treatment strategies.

About 90% of follicular lymphomas harbour the t(14;18) translocation, which results in overexpression of BCL2.[7] Recent sequencing studies have catalogued additional recurrent genetic alterations in follicular lymphomas,[7] but their effects on clinical outcome remain unclear. We postulated that integrating gene mutations into prognostic models will improve risk stratification for patients

with follicular lymphoma receiving first-line immunochemotherapy. To address this hypothesis, we analysed the full coding sequence of 74 genes that are recurrently mutated in lymphomas across two independent cohorts of patients who received first-line immunochemotherapy.

Methods

Study design and participants

We did a retrospective analysis of gene mutations and clinical risk factors in two cohorts of patients with symptomatic, advanced stage, or bulky follicular lymphoma grade 1, 2, or 3A, who had a biopsy specimen obtained 12 months or less before initiation of a rituximab-containing first-line immunochemotherapy regimen.

The training cohort (figure 1) was derived from the randomised GLSG2000 trial[8] of the German Low-Grade Lymphoma Study Group (GLSG). Between May 4, 2000, and Oct 20, 2010, this trial recruited patients with a diagnosis of stage III/IV follicular lymphoma. All patients needed treatment, as defined by the presence of B symptoms, bulky disease (mediastinal lymphomas >7·5 cm or other lymphomas >5 cm), impairment of normal haemopoiesis, or rapidly progressive disease.[8] Patients eligible for potentially curative radiotherapy were excluded. Other exclusion criteria were age younger than 18 years, pregnancy, Eastern Cooperative Oncology Group (ECOG) performance status 3 or more, or severe organ dysfunction unless due to lymphoma. Patients with other systemic malignancies, previous radiotherapy, chemotherapy, or immunotherapy were ineligible.[8]

Patients were excluded from the training cohort if the diagnosis of follicular lymphoma was not confirmed by central pathology review, or if they were randomised to not receive rituximab, or randomised to receive post-remission high-dose radiochemotherapy followed by autologous stem-cell transplantation (figure 1) because these regimens do not represent current standard practice.[2] The remaining patients received rituximab 375 mg/m², plus cyclophosphamide 750 mg/m², doxorubicin 50 mg/m², and vincristine 1·4 mg/m² (maximum, 2·0 mg) on day 1, and prednisone 100 mg/m² on days 1–5 (R-CHOP), every 3 weeks for a total of eight cycles.[8] Patients achieving at least a partial response after six cycles received two additional cycles of R-CHOP followed by interferon α maintenance.[8] This trial was approved by the institutional review board and patients signed informed consent that included molecular and genetic analyses.

The validation cohort (figure 1) consisted of patients with symptomatic, advanced stage, or bulky follicular lymphoma considered ineligible for curative irradiation, receiving first-line immunochemotherapy with six to eight cycles of R-CVP (rituximab 375 mg/m² plus cyclophosphamide 1000 mg/m², and vincristine 1·4 mg/m² on day 1, and prednisone 100 mg/day, days 1–5) every 3 weeks at the British Columbia Cancer Agency (BCCA). Diagnostic tumour biopsies were collected between Feb 24, 2004, and Nov 24, 2009, as part of a research project approved by the University of British Columbia–BCCA Research Ethics Board that included molecular and genetic analyses. From 2006 onwards, patients achieving at least a partial response were scheduled for rituximab maintenance (375 mg/m² given every 3 months for a total of eight doses; table 1).

Procedures

We analysed the full coding sequence of 74 genes selected based on the presence of recurrent mutations (appendix pp 7, 8). Genomic DNA from formalin-fixed paraffin-embedded pretreatment tumour biopsies was used for customised hybrid-capture target enrichment (SureSelect, Agilent, Santa Clara, CA, USA) and Illumina sequencing, as previously described.[9] Non-tumour DNA from peripheral blood or bone marrow samples was sequenced to exclude germ-line polymorphisms and platform-specific artifacts (appendix p 2). Matched non-tumour samples were obtained from 24 GLSG2000 patients who tested negative or had less than 1% minimal residual disease after being treated, as previously described,[10] and for ten BCCA patients free of lymphoma cell infiltration by routine clinical assessment, allowing for the filtering of somatic mutations. A panel-of-normals filter was generated from all matched (34) and other (25) non-tumour controls (appendix p 2) available for this platform, as previously described.[11] Tumour samples were filtered using the panel-of-normals and variants were rejected as germ-line events or sequencing artifacts if present in two or more normal samples. Known germ-line polymorphisms from the Exome Sequencing Project and the dbSNP (build 142) databases were also excluded. For all genes except for *BCL2*, we analysed only non-silent mutations (missense mutations, nonsense mutations, in-frame or frame-shift insertions or deletions [InDel], translational start site mutations, splice site mutations) with variant allele frequencies of 10% or more. For *BCL2*, any variant with variant allele frequencies of 10% or more was reported as hyper-mutation. 100% (46/46, appendix p 9) of a randomly selected set of mutations were validated by Sanger sequencing. We applied the MutSigCV algorithm[12] to identify genes that were mutated more often than expected by chance given background mutation processes. We used the evolutionary conservation of the affected aminoacid in protein homologues[13] to predict the effect of non-silent mutations.

See Online for appendix

For the **Exome Sequencing Project** see http://evs.gs.washington.edu/EVS/

Statistical analysis

For the training set, we did univariable and multivariable analyses for all genes mutated in five or more patients for stability reasons of the models, as previously performed.[14] Genes were studied for their correlation with the binary FLIPI (high-risk *vs* low or intermediate-risk) and its individual components, ECOG performance status and sex. We used the binary FLIPI because no significant

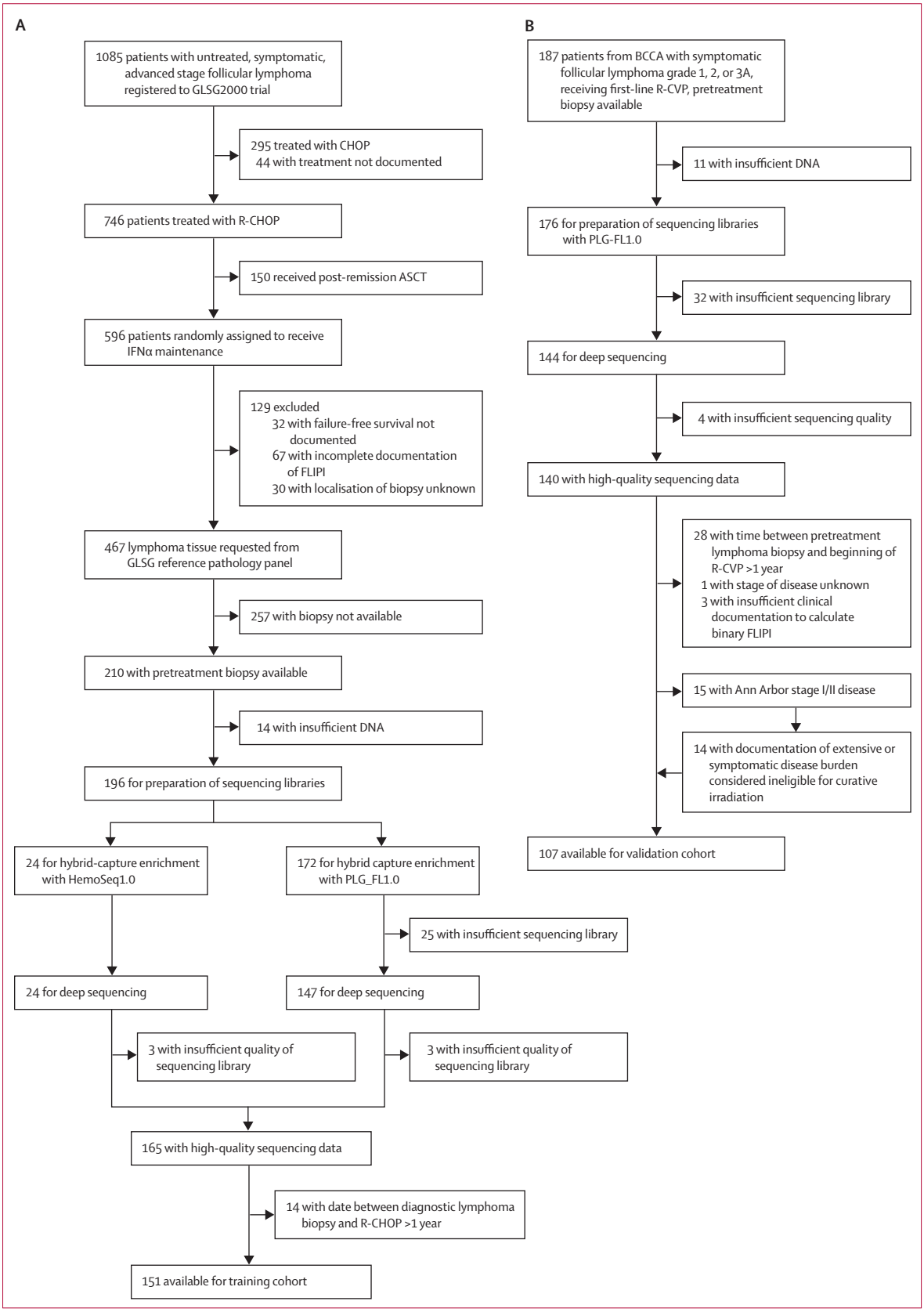

Figure 1: **Patient flow for the GLSG2000 training cohort (A) and BCCA validation cohort (B)**
HemoSeq1.0 and PLG_FL1.0 are the exon capture strategies (appendix pp 7, 8). BCCA=British Columbia Cancer Agency. ASCT=autologous stem-cell transplantation. IFNα=interferon α. FLIPI=Follicular Lymphoma International Prognostic Index. GLSG=German Low-Grade Lymphoma Study Group.

differences for failure-free survival were seen between low-risk and intermediate-risk patients treated with R-CHOP in a previous study,[6] and a recent update of this analysis with a median follow-up of 7·8 years (appendix p 13). Univariable Cox regression was done with and without adjustment for binary FLIPI and ECOG performance status. The Bonferroni-Holm procedure was used to adjust for multiple testing.

We generated risk models for failure-free survival using different sets of predictors. One consisted only of recurrent gene mutations and the second also included the binary clinical variables FLIPI and ECOG performance status. The latter were chosen because they were confirmed to be independent risk factors for failure-free survival in a large cohort of assessable patients from the GLSG2000 trial who received R-CHOP (appendix p 13). Additionally, we calculated prognostic models with selected genes grouped into well-described functional pathways (appendix p 5). L1-penalised Cox regression (Lasso) with 10-fold cross-validation was used to select gene mutation predictors. To generate prognostic risk models for failure-free survival, we used all 31 genes with mutations in five or more patients irrespective of results from univariable testing. We postulated that clinical variables might add prognostic information beyond tumour biology, and calculated risk models with and without inclusion of the previously well-established risk factors, FLIPI[5] and ECOG performance status.[5,15,16] We favoured FLIPI for inclusions in the model over the previously described FLIPI-2[17] because β-2-microglobulin serum concentrations are not routinely measured, particularly in North America, and the original FLIPI is more commonly used.[18] β-2-microglobulin serum concentrations were only available from a subset of patients from the GLSG2000 trial (84/151) and not for BCCA patients. We also calculated the FLIPI-2,[17] which is comprised of elevated β-2-microglobulin, longest diameter of largest node greater than 6 cm, bone marrow involvement, haemoglobin less than 120 g/dL, and age older than 60 years, for 126 assessable patients with available data from the GSLG2000 cohort.

For the model that included FLIPI and ECOG performance status, these categorical variables were not penalised and thus forced into the model. A prognostic score was calculated as the sum of predictor values weighted by Lasso coefficients. The cutoff value maximising the log-rank statistics dichotomised patients into high-risk and low-risk groups. We used bootstrap on the training cohort and selected the model with the largest bootstrap-corrected hazard ratio (HR) for failure-free survival of high-risk versus low-risk groups. Outcome data for the validation cohort remained masked until final validation. Sensitivity and specificity of the final model were estimated with the inverse probability of censoring weighting (IPCW) approach, as previously described.[19] Additionally, we calculated the *C*-index (Harrell) for prognostic discrimination.

	GLSG2000 training cohort	BCCA validation cohort	p value
Patients			
Number of assessable patients*	151	107	
Male	78 (52%)	59 (55%)	0·67
Female	73 (48%)	48 (45%)	
Clinical risk factors			
>60 years	57 (38%)	59 (55%)	0·0083
>4 nodal sites	106 (70%)	78 (73%)	0·74
Lactate dehydrogenase elevated	49 (32%)	22 (21%)	0·074
Haemoglobin <120 g/L	32 (21%)	12 (11%)	0·062
ECOG performance status >1	8 (5%)	16 (15%)	0·016
FLIPI high-risk	77 (51%)	53 (50%)	0·92
Treatment			
First-line treatment	R-CHOP†	R-CVP‡	
Maintenance treatment	IFNα	Rituximab	
Number of patients intended for maintenance treatment	151	93	
Outcome			
5-year failure-free survival (95% CI; number of events)	66·22% (58·63–74·79; 63)	58·43% (49·73–68·66; 48)	
5-year overall survival (95% CI; number of deaths)	83·25% (77·20–89·78; 33)	74·40% (66·50–83·23; 32)	
Median (IQR) follow-up for overall survival, years	7·7 (5·5–9·3)	6·7 (5·7–7·6)	

ECOG=Eastern Cooperative Oncology Group. FLIPI=Follicular Lymphoma International Prognostic Index. *Assessable patients fulfilled all of the following inclusion criteria: confirmed diagnosis of follicular lymphoma by reference pathology, symptomatic disease requiring treatment, advanced stage disease or bulky disease considered ineligible for curative irradiation, initial treatment regimen contained rituximab, no dose-intensified consolidative treatment, diagnostic lymphoma tissue or DNA available, time between pretreatment biopsy and treatment initiation less than 1 year, and high-quality sequencing data available. †R-CHOP: eight cycles of rituximab 375 mg/m², cyclophosphamide 750 mg/m², doxorubicin 50 mg/m², vincristine 1·4 mg/m² (day 1), and prednisone 100 mg (days 1–5), every 3 weeks, no consolidative transplant; patients achieving at least a partial response received interferon α (IFNα) maintenance (3–5 MioU per week). ‡R-CVP: six to eight cycles of rituximab 375 mg/m², cyclophosphamide 1000 mg/m², vincristine 1·4 mg/m² (day 1), and prednisone 100 mg/day (days 1–5), every 3 weeks; from 2006 onwards, rituximab maintenance (375 mg/m² given every 3 months for a total of eight doses) was given to patients achieving at least a partial response.

Table 1: **Patient and disease characteristics**

All statistical analyses were done using the statistical software R (version 3.1.0) and the R-packages ggplot2 (version 1.0.1), corrplot (version 0.73), MEMo (version 1.0), survival (version 2.38-1), timeROC (version 0.3), penalised (version 0.9-42), pec (version 2.4.4), and SAS 9.2.

Additional methods are described in the appendix (pp 2–4).

Role of the funding source

The funding source had no involvement in study design, collection, analysis, and interpretation of the data, and in writing of the report or decision to submit this paper for publication. AP, RK, MM, PVH, MDu, HPS, RDG, DMW, and OW had access to the raw sequencing data. VJ, EH, MU, and OW had access to raw clinical data of GLSG2000 patients. VJ, RK, EH, AAM, MU, RDG, and OW had access to raw clinical data of BCCA patients. The corresponding author had full access to all the data in the study and the final responsibility for the decision to submit for publication.

Results

The training cohort consisted of 151 GLSG2000 patients (table 1, figure 1). The median age was 57 years (IQR 50–64), 78 (52%) were male, and 77 (51%) had a high-risk FLIPI. After a median follow-up of 7·7 years (IQR 5·5–9·3), 5-year failure-free survival was 66·22% (95% CI 58·63–74·79; 63 events overall) and 5-year overall survival was 83·25% (95% CI 77·20–89·78; 33 deaths overall; table 1).

The median number of targeted genes with non-silent mutations per patient was four (IQR 3–5·5; appendix p 14). Nine genes had non-silent mutations in 10% or more of patients (*KMT2D* 79% [n=119], *CREBBP* 70% [105], *TNFRSF14* 32% [48], *EZH2* 22% [33], *ARID1A* 15% [22], *EP300* 14% [21], *CARD11* 12% [18], *STAT6* 11% [16], and *MEF2B* 10% [15]). *BCL2* was hypermutated in 76% (115) of patients. Mutation details on all significantly mutated genes[12] and on eight additional genes with non-silent mutations in more than 5% of patients are shown in the appendix (p 15).

Overall, 146 (97%) follicular lymphomas harboured non-silent mutations in epigenetic modifiers, with predominantly disruptive mutations found in *KMT2D*, *CREBBP*, *EP300*, *ARID1A*, and *BCL7A*. Additional disruptive mutations affected *TNFRSF14*, the B-cell receptor components *CD79A* and *CD79B*, the NFκB pathway inhibitor *TNFAIP3*, and the transcription factors *IRF8* and *ETS1*. Overall, 70 follicular lymphomas (46%) harboured one or more non-silent mutation in lymphoid transcription factors (appendix p 15).

A subset of mutations clustered at known hotspots. Mutations in EZH2, the catalytic subunit of the polycomb repressor complex 2 (PRC2), were clustered at Tyr641 (29 [88%] of 33 patients) and Ala677 (three [9%] of 33 patients).[20,21] Other hotspot and clustered mutations affected CARD11 (16 [80%] of 20 in the coiled-coil domain),[22] CXCR4 (WHIM-like mutations in five [100%] of five),[23] exon 1 of *FOXO1* (nine [90%] of 10),[24] STAT6 at Asp419 (seven [44%] of 16),[25] and MEF2B mutations (15 [100%] of 15 within the N-terminal domain;[26] appendix p 16). By contrast, *PIM1* and *BCL2*, which are known targets of aberrant somatic hypermutation,[27] predominantly harboured transition mutations (appendix p 17) with low predicted functional effect. Additional analyses for co-occurring mutations are included in the appendix (pp 5, 18).

The FLIPI and many individual FLIPI components were associated with specific gene mutations in univariable analyses (appendix p 10). However, none of these associations were statistically significant after correction for multiple testing. In univariable analyses, mutations in *EP300* (HR 1·99, 95% CI 1·08–3·68; p=0·028) and *FOXO1* (HR 2·74, 1·23–6·09; p=0·013) were associated with shorter failure-free survival, whereas mutations in *EZH2* (HR 0·46, 0·22–0·93; p=0·030) were associated with longer failure-free survival. Adjustment for FLIPI and ECOG performance status also revealed that *ARID1A* mutations correlated with longer failure-free survival (HR 0·40, 95% CI 0·16–0·996; p=0·049) and *TP53* mutations with shorter failure-free survival (HR 2·85, 1·12–7·27; p=0·029; appendix p 19). After correction for multiple testing by Bonferroni-Holm, no single mutation (including *BCL2* hypermutation) was significantly associated with failure-free survival (appendix pp 19, 20).

TP53 mutations were associated with inferior overall survival as a single factor (HR 4·70, 95% CI 1·80–12·23; p=0·0015) and with adjustment for FLIPI and ECOG performance status (HR 6·32, 2·35–17·01; p=0·00026). *CARD11* mutations were prognostic for inferior overall survival in univariable analysis (HR 2·43, 95% CI 1·09–5·39; p=0·029) and when adjusted for FLIPI and ECOG performance status (HR 3·71, 1·59–8·67; p=0·0024). After correction for multiple testing, only *TP53* mutations remained significantly associated with overall survival (adjusted p=0·0081; appendix p 19).

We generated multivariable risk models for failure-free survival using different sets of predictors: one consisted only of recurrent gene mutations, and the second also included the binary clinical variables FLIPI and ECOG performance status. Additional risk models are described in the appendix (pp 5, 12). Internal validation by bootstrap analysis showed superiority of the model that integrated gene mutations and clinical factors compared to the model of only gene mutations and the other models (appendix p 12). This clinicogenetic model, which we termed m7-FLIPI, was calculated as the sum of predictor values weighted by Lasso coefficients, and included high-risk FLIPI (β_{Lasso}=+0·79), poor ECOG performance status (>1, β_{Lasso}=+0·38), and non-silent mutations in seven genes: *EZH2* (β_{Lasso} –0·53), *ARID1A* (–0·4), *EP300* (+0·33), *FOXO1* (+0·26), *MEF2B* (–0·07), *CREBBP* (+0·05), and *CARD11* (+0·04; figure 2). To divide the risk score into high-risk and low-risk cohorts, a cutoff of 0·8 was calculated to be optimum (appendix p 3). The m7-FLIPI identified a high-risk group (43 [28%] of 151 patients) with 5-year failure-free survival of 38·29% (95% CI 25·31–57·95) and a low-risk group (108 [72%] of 151 patients) with 5-year failure-free survival of 77·21% (69·21–86·14 [HR 4·14, 2·47–6·93; p<0·0001; bootstrap-corrected HR 2·02]), and outperformed a prognostic model of only gene mutations (HR 3·76, 2·10–6·74; p<0·0001; bootstrap-corrected HR 1·57). In the patients with available FLIPI-2 scores (n=126), the m7-FLIPI (p<0·0001) outperformed the FLIPI-2 (p=0·00088; appendix p 21).

The validation cohort consisted of 107 BCCA patients (figure 1). Median age was 62 years (IQR 54–69) and 59 (55%) were male. 93 (87%) of 107 patients received rituximab maintenance. Compared with the GLSG2000 cohort, the BCCA cohort included more patients older than 60 years and more with ECOG performance status greater than 1, whereas elevated lactate dehydrogenase and haemoglobin less than 120 g/L were less frequent

(table 1). The fraction of high-risk FLIPI patients was similar between the two cohorts (53 [50%] of 107 *vs* 77 [51%] of 151; p=0·92). After a median follow-up of 6·7 years (IQR 5·7–7·6) in the BCCA cohort the 5-year failure-free survival was 58·43% (95% CI 49·73–68·66) and the overall survival was 74·40% (66·50–83·23).

The median number of mutations in the validation cohort was five (IQR 3–6; appendix p 14). The targeted mutational landscape of the validation cohort is summarised in the appendix (p 22). Compared with the training cohort, there were no significant differences in the mutation frequencies of any of the 74 genes after correction for multiple testing. m7-FLIPI defined high-risk (24 [22%] of 107 patients) and low-risk (83 [78%] of 107 patients) groups with 5-year failure-free survival of 25·00% (95% CI 12·50–49·99) and 68·24% (58·84–79·15), repectively (HR 3·58, 2·00–6·42; p<0·0001; figure 2, table 2, appendix p 11). The m7-FLIPI validated and outperformed FLIPI alone (HR 2·18, 95% CI 1·21–3·92), and FLIPI combined with ECOG performance status (HR 2·03, 95% CI 1·12–3·67); figure 2, appendix p 20). Performance metrics for m7-FLIPI and FLIPI based on 5-year failure-free survival in both cohorts are outlined in table 2.

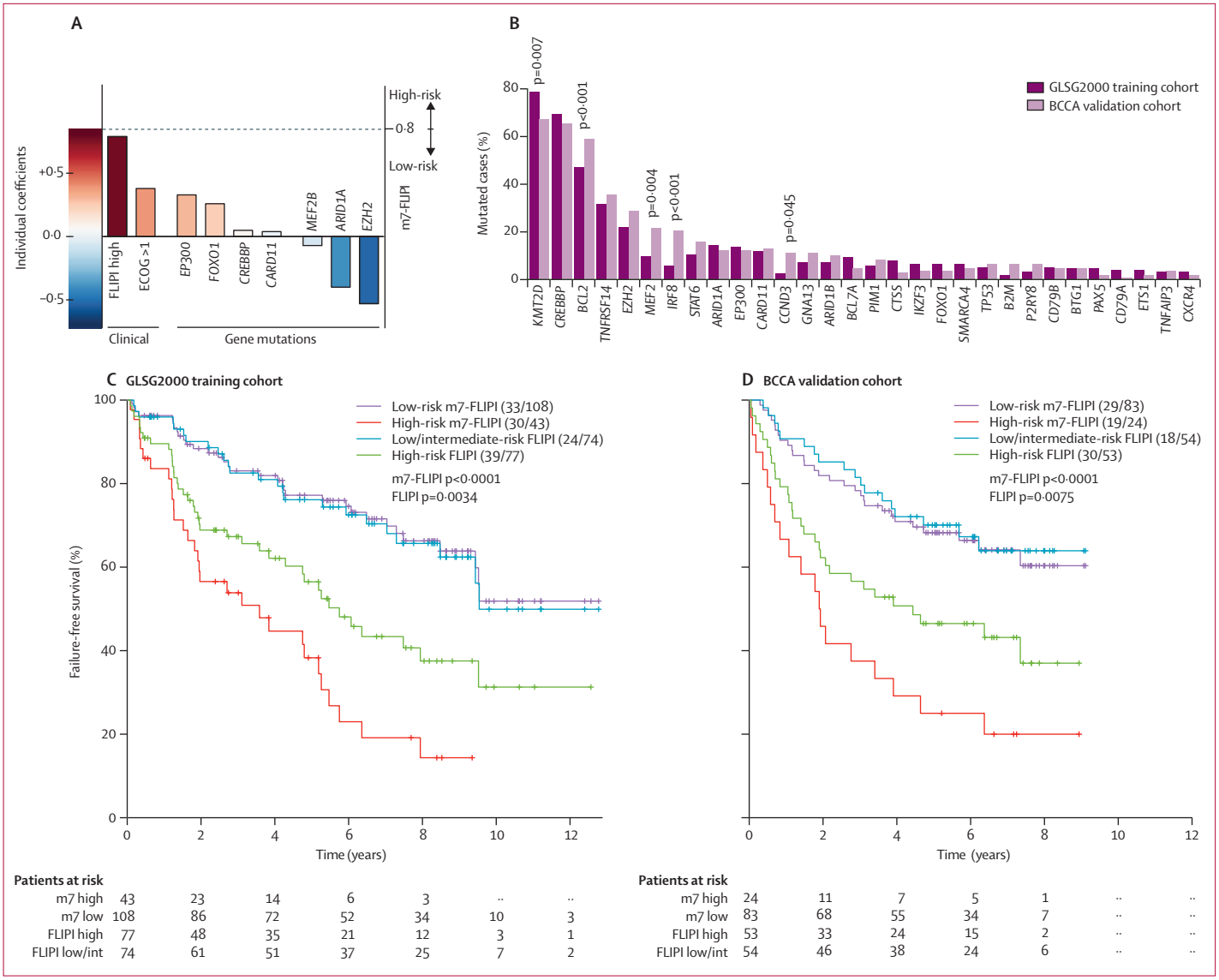

Figure 2: **The clinicogenetic risk model m7-FLIPI**
(A) The m7-FLIPI (m7) is calculated as the sum of individual clinical and gene mutation predictor values weighted by their individual coefficients. (B) Mutation frequencies of the GLSG2000 training and the BCCA validation cohorts. p values by Fisher's exact test, without correction for multiple testing. Depicted are all significantly mutated genes[12] and genes with non-silent mutations in more than 5% of cases from the GLSG2000 training cohort. Detailed mutation plots for both cohorts are shown in the appendix (pp 15, 22). (C) Kaplan-Meier curves for failure-free survival for the GLSG2000 training cohort by FLIPI and by m7-FLIPI. (D) Kaplan-Meier curves for failure-free survival for the BCCA validation cohort by FLIPI and by m7-FLIPI. Numbers in parentheses show number of patients with event/number of patients per cohort. FLIPI low/int=low or intermediate-risk FLIPI.

	High-risk patients, % (n/N)	High-risk *vs* non-high-risk 5-year failure-free survival (95% CI)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	*C* index (95% CI)
GLSG2000 cohort m7-FLIPI	28% (43/151)	38·29% (25·31–57·95) *vs* 77·21% (69·21–86·14)	52%	85%	64%	78%	0·80 (0·71–0·89)
BCCA cohort m7-FLIPI	22% (24/107)	25·00% (12·50–49·99) *vs* 68·24% (58·84–79·15)	41%	89%	72%	68%	0·79 (0·69–0·89)
GLSG2000 cohort FLIPI	51% (77/151)	56·47% (45·61–69·92) *vs* 76·14% (66·54–87·12)	65%	60%	45%	77%	0·70 (0·58–0·82)
BCCA cohort FLIPI	50% (53/107)	46·49% (34·69–62·30) *vs* 70·08% (58·81–83·52)	64%	62%	55%	71%	0·70 (0·58–0·83)

Sensitivity, specificity, positive predictive value, and negative predictive value for patients remaining without failure at 5 years. Harrell's *C* index or concordance *C* is a generalisation of the area under the receiver operating characteristic curve for survival data and quantifies prognostic discrimination. FLIPI=Follicular Lymphoma International Prognostic Index.

Table 2: **Performance metrics for the m7-FLIPI and FLIPI based on 5-year failure-free survival**

Although m7-FLIPI was developed to predict failure-free survival, we tested its prognostic utility for overall survival. High-risk m7-FLIPI was associated with an inferior 5-year overall survival of 65·25% (95% CI 51·40–82·84) versus 89·98% (84·26–96·08; p=0·00031) in the training cohort, and 41·67% (25·95–66·89) versus 84·01% (76·39–92·39; p<0·0001) in the validation cohort (appendix p 21). In each cohort, the m7-FLIPI again outperformed the FLIPI alone (appendix p 23).

Distinct molecular features exist between t(14;18)-positive and t(14;18)-negative follicular lymphomas.[28] Fluorescence-in-situ hybridisation (FISH) data using a BCL2 break-apart probe (appendix p 4) was available for 97 GLSG2000 patients and 104 BCCA patients. t(14;18) was present in 88 (91%) assessable GLSG2000 patients and 92 (88%) assessable BCCA patients. Similar to the overall study population, m7-FLIPI outperformed the FLIPI and was significantly associated with failure-free survival in the subset of t(14;18)-positive cases from both cohorts (appendix p 24). Because of the small number of t(14;18)-negative cases in each cohort, it was not feasible to test the prognostic utility of m7-FLIPI specifically within this population.

In both cohorts, the improved performance with m7-FLIPI resulted from reclassification of a subset of patients with high-risk FLIPI into the low-risk m7-FLIPI category (figure 3). 34 (44%) of 77 patients in the training cohort and 29 (55%) of 53 patients in the validation cohort that were classified as high-risk by FLIPI were re-classified into the low-risk m7-FLIPI group (figure 3). Tumours from these patients were enriched for mutations in *EZH2* (50% [17/34] *vs* 0% [0/43] and 55% [16/29] *vs* 0% [0/24]), *MEF2B* (29% [10/34] *vs* 0% [0/43] and 28% [8/29] *vs* 13% [3/24]), and *ARID1A* (32% [11/34] *vs* 2% [1/43] and 31% [9/29] *vs* 4% [1/24]). By contrast, samples from patients classified as high-risk by m7-FLIPI were enriched for mutations in *EP300* and *CREBBP* in both the training and validation cohorts (figure 3, appendix p 25).

The finding that all 66 patients across both cohorts with *EZH2* mutations were classified as low-risk m7-FLIPI led us to ask whether *EZH2* expression defines a unique biology in follicular lymphoma. To address this, we profiled gene expression in the 107 BCCA patients from the validation cohort and 33 additional BCCA patients that were excluded from the validation cohort for stringency reasons (28 patients, time between lymphoma biopsy and initiation of R-CVP >1 year; five patients, insufficient clinical documentation; figure 1). Whole-genome gene expression profiling using the Illumina cDNA-mediated annealing, selection, extension, and ligation assay (DASL; appendix p 4) was done for 140 BCCA tumour samples with available high-quality sequencing data (figure 1) and successful for 138 cases, including 106 patients from the validation cohort.

All assessable tumour samples were divided into mutated and non-mutated cases for the 15 most common gene mutations and compared for differentially expressed genes (p<0·05; appendix p 26). A false discovery rate approach to adjust individual p values revealed that *EZH2* was associated with the highest number (129) of differentially expressed genes at q values less than 0·05 (appendix p 26). These genes were used to define a distinct gene expression signature for *EZH2* mutation status by unsupervised clustering of the 106 patients from the validation cohort (appendix p 27). This signature significantly correlated with both failure-free survival and overall survival (appendix p 28). Gene set enrichment analysis of our *EZH2* signature showed significant enrichment of a previously reported *EZH2* signature[29] (figure 4).

To measure how well this gene expression signature correctly identified the presence or absence of *EZH2* mutation, we calculated the accuracy (ie, the proportion of true positives and true negatives in the population). Overall, the *EZH2* mutation status was correctly allocated by this gene expression signature in 93 (88%) of 106 tumour samples, indicating that the mutation is truly associated with a distinct transcriptional profile. The presence of *EZH2* mutations was associated with a significant improvement in both failure-free survival

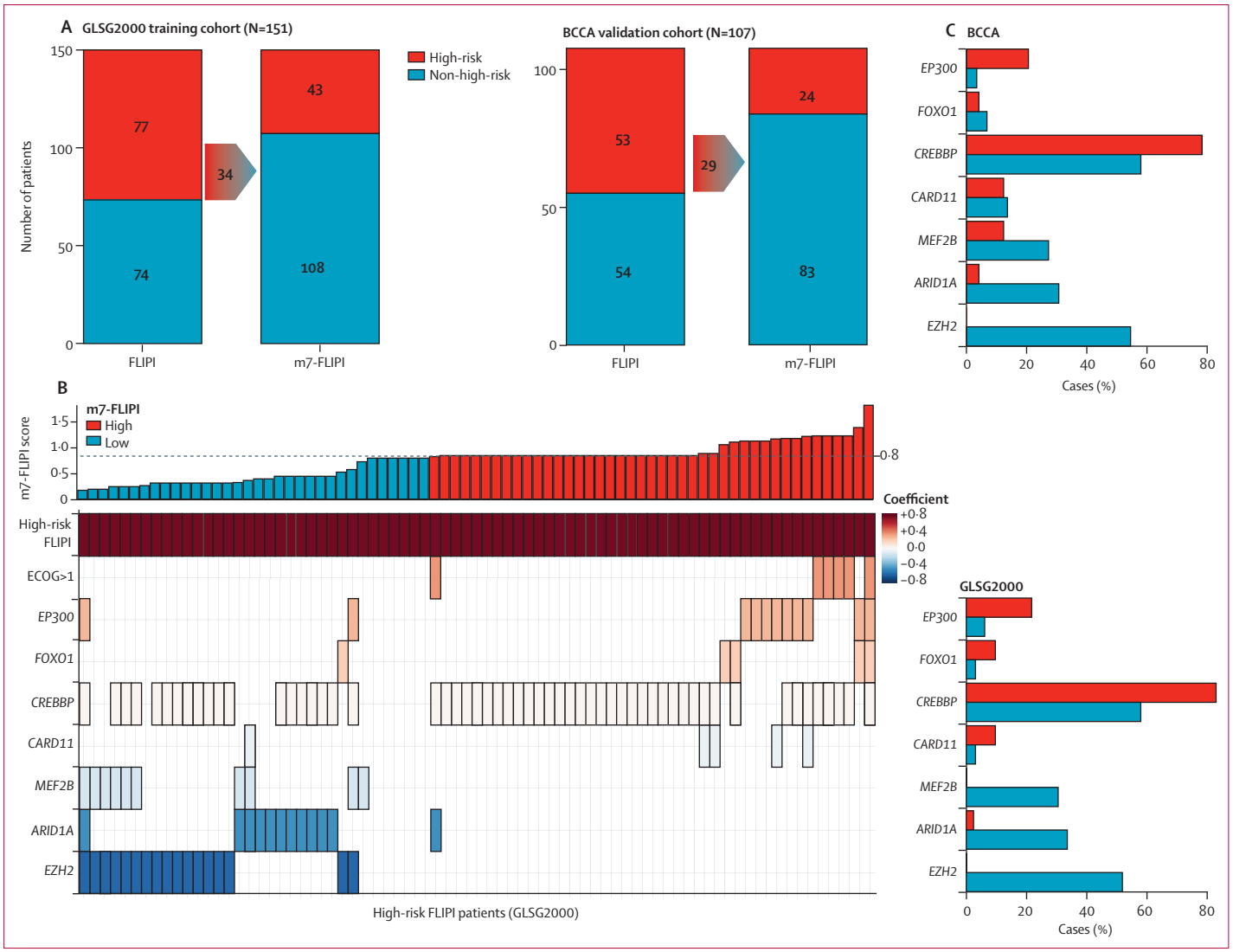

Figure 3: **Reclassification of risk category by m7-FLIPI**

(A) Migration plot showing reclassification of patients by m7-FLIPI in both cohorts. (B) m7-FLIPI score for all high-risk FLIPI patients from the GLSG2000 cohort, along with the Eastern Cooperative Oncology Group performance status (ECOG) and molecular predictors. Boxes indicate high-risk FLIPI, an ECOG performance status of more than 1, or a mutation in the indicated gene, and the colour code indicates the coefficient of the individual m7-FLIPI predictor. The corresponding Kaplan-Meier curves for failure-free survival and overall survival for patients classified as high-risk by FLIPI and reclassified as high-risk versus low-risk by m7-FLIPI are shown in the appendix (pp 30, 31). Results from the BCCA cohort are shown in the appendix (p 25). On the right, relative frequencies of molecular predictors by m7-FLIPI category in high-risk FLIPI patients from the GLSG2000 cohort are shown. (C) Relative frequencies of molecular predictors by m7-FLIPI category in high-risk FLIPI patients from the BCCA cohort.

(figure 4) and overall survival (appendix p 29) in both cohorts, but only among patients with high-risk FLIPI.

Discussion

To our knowledge, we report the largest study of recurrent and significant mutations in patients with symptomatic follicular lymphoma who received first-line immunochemotherapy and the first to assess prognostic relevance of mature data from two independent cohorts. By adding the mutational status for seven genes to established clinical risk factors, we developed an improved prognostic algorithm to help clarify which patients are likely to have poor outcome after standard immunochemotherapy. Our cohorts span two continents and include both a clinical trial population (GLSG2000) and a population-based registry (BCCA). We believe the fact that the validation cohort received a different immunochemotherapy and maintenance regimen and gave virtually the same results supports the broad applicability of the m7-FLIPI.

By contrast with previous studies that focused primarily on single gene alterations,[30–33] we took a multivariable approach that included a comprehensive compilation of recurrent gene mutations and clinical

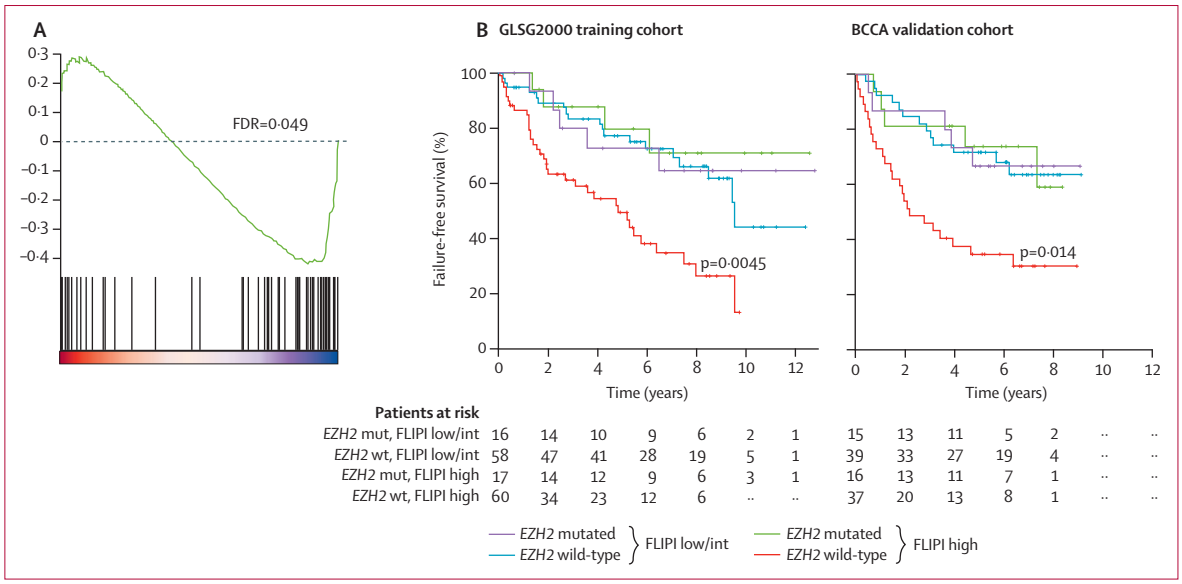

Figure 4: **Biological and prognostic relevance of *EZH2* mutations in follicular lymphoma**
(A) Gene set enrichment analysis of the *EZH2* signature from an independent cohort in our *EZH2* signature.[29] (B) Effect of *EZH2* mutation status combined with binary FLIPI risk score on failure-free survival. p values are shown for the comparison of patients classified as high-risk by FLIPI and do not have *EZH2* mutations versus patients that are classified as high-risk by FLIPI and have *EZH2* mutations. FDR=false discovery rate. low/int=low or intermediate. wt=wild-type. mut=mutant.

risk factors. The advantage of multivariable modelling is that it is unbiased by biological assumptions, and thereby reflects the fact that interactions between distinct gene mutations and clinical factors are complex, interdependent, and largely unknown. This approach does not require that single gene mutations have a significant effect on outcome. In fact, gene mutations significantly associated with outcome by univariable analysis (eg, in *TP53*) can drop out. The superior performance of the m7-FLIPI shows that both clinical factors that reflect the patient's performance status and extent of disease as well as gene mutations affect treatment outcome, and should be combined to provide optimum prognostic information.

In both cohorts, about half the patients classified as high-risk using FLIPI were classified as low-risk using m7-FLIPI and these patients had outcomes indistinguishable from those with low-risk FLIPI. This reclassification of risk category using the m7-FLIPI score primarily results from gene mutations, in particular mutation of *EZH2*, that lower the risk of a failure-free survival event after immunochemotherapy. By contrast, the remaining high-risk patients were enriched for poor outcome, with 5-year failure-free survival in the GLSG2000 cohort of only 38·29% (95% CI 25·31–57·95) and 25·00% (12·50–49·99) in the BCCA cohort.

Although m7-FLIPI was developed for failure-free survival, it was also prognostic for overall survival. However, low-risk m7-FLIPI does not necessarily indicate a more indolent disease course, as all patients studied had required treatment. We did not study asymptomatic patients who did not require therapy; doing so is particularly challenging because lead time (ie, the time between diagnosis and symptomatic disease requiring treatment) depends on many variables other than disease biology. It is also important to note that all patients in both cohorts had biopsies obtained within 12 months before beginning treatment. The genetics of untreated follicular lymphoma might change within a patient over the course of time, so it remains unclear whether m7-FLIPI is applicable to patients whose sequenced biopsy was obtained many years before receiving first-line immunochemotherapy.

A previous study reported improved prediction of failure-free survival using the FLIPI-2 score.[17] However, we were only able to compare m7-FLIPI with FLIPI-2 in a subset of GLSG2000 patients; thus, it will be important to further validate m7-FLIPI in additional cohorts that have complete documentation of all FLIPI-2 variables. Additionally, the predictive value of m7-FLIPI will need to be assessed in patients who receive treatment regimens containing other chemotherapeutics (eg, bendamustine) or alternative anti-CD20-directed antibodies. Further studies are needed in patients with t(14;18)-negative follicular lymphoma, as the small numbers of t(14;18)-negative cases in our cohorts precluded an adequately powered assessment of the prognostic utility specifically within this rare population. Studies are also needed to determine whether mutations present at less than 10% variant allele frequency can further guide prognostication. Finally, with evolving omics and other technologies, future studies will be needed to iteratively improve m7-FLIPI by adding or substituting genetic, epigenetic, proteomic, or other factors.

Many B-cell lymphomas, both with and without activating *EZH2* mutations, might depend on EZH2 function.[34] As a result, inhibitors of EZH2 activity are in clinical trials for patients with relapsed and refractory lymphomas.[34] Mutations in *EZH2* affected 33 (22%) of 151 GLSG200 patients and 31 (29%) of 107 BCCA patients and clustered at the Tyr641 and Ala677 hotspots that are known to promote hypertrimethylation of lysine 27 on histone H3 (H3K27)[20,21] and B-cell transformation.[35] Unexpectedly, these mutations were strongly associated with low-risk m7-FLIPI, improved failure-free survival and overall survival, and defined a unique transcriptional signature. Thus, patients who harbour *EZH2* mutations are likely to have good outcomes after conventional immunochemotherapy.

In summary, by adding the mutational status for seven genes to established clinical risk factors, we developed an improved prognostic algorithm that can be applied to patients receiving first-line immunochemotherapy. A freely accessible online tool is now available to calculate the m7-FLIPI. If the m7-FLIPI is further validated in subsequent studies, it could serve as a valuable biomarker to select patients for trials of risk-adapted treatment strategies (eg, dose intensification or novel molecular targeted agents).

For the **online tool** see http://www.glsg.de/m7-FLIPI

Contributors

AP and VJ contributed to data analysis, data interpretation, and figures. RK provided patient samples, did experiments, and contributed to data collection, data analysis, data interpretation, and figures. EH contributed to study design, data analysis, data interpretation, and figures. AMS, MS, CP, HH, and DE provided patient samples, and contributed to data collection. NK, AM, and AS did experiments. MM and AAM contributed to data collection. EL did experiments and contributed to data analysis. PVH, MDu, HPS, and CH contributed to data analysis. JMC and LHS collected patient data and contributed to data interpretation. MDr collected patient data. DN contributed to study design, data analysis, and data interpretation. PM, ACF, MLH, HS, AR, GO, and WK provided patient samples. MU contributed to data collection, data analysis, and data interpretation. WH contributed to data collection and data interpretation. RDG provided patient samples, collected patient data, and contributed to study design, data interpretation, and manuscript writing. DMW contributed to study design, data analysis, data interpretation, and manuscript writing. OW did experiments, collected patient data, and contributed to study design, data analysis, data interpretation, figures, and manuscript writing.

Declaration of interests

EH reports travel support from Roche Pharma AG, outside the submitted work. MM reports grants from National Cancer Institute (institutional research training grant T32 CA009172, awarded to the Department of Medical Oncology at the Dana-Farber Cancer Institute), during the conduct of the study. LHS reports honoraria from from Roche/Genentech, Lundbeck, Celgene, Seattle genetics, Janssen, and Amgen, outside the submitted work. WK reports grants from Roche, Celgene, Novartis, and Janssen, outside the submitted work. RDG reports personal fees from Genentech, Seattle Genetics, Celgene, and Janssen, outside the submitted work. DMW reports grants and personal fees from Novartis, outside the submitted work. The other authors declare no competing interests.

Acknowledgments

This research was supported by the Max-Eder Program of the Deutsche Krebshilfe e.V. (number 110659 to OW) and in part by a Program Project Grant from the Terry Fox Research Institute to JMC and RDG (number 1023). AP is a Mildred-Scheel Postdoctoral Research Fellow of the Deutsche Krebshilfe e.V. (number 111354). RK is supported by fellowships from the Ligue Genevoise Contre le Cancer et Fondation Dr Henri Dubois-Ferrière Dinu Lipatti, the Canadian Institutes for Health Research, the Michael Smith Foundation for Health Research, and the University of British Columbia. AM holds a fellowship awarded by the Mildred Scheel Cancer Foundation of the Deutsche Krebshilfe e.V. CH is funded by a postdoctoral fellowship from The Alfred Benzon Foundation. JMC is partially supported by Genome British Columbia, Genome Canada, the Canadian Institutes for Health Research, and the British Columbia Cancer Foundation. DMW is a Leukemia and Lymphoma Society Scholar. The GLSG2000 trial was partially supported by Roche Pharma AG.

References

1 Anderson JR, Armitage JO, Weisenburger DD. Epidemiology of the non-Hodgkin's lymphomas: distributions of the major subtypes differ by geographic locations. Non-Hodgkin's Lymphoma Classification Project. *Ann Oncol* 1998; **9:** 717–20.

2 Hiddemann W, Cheson BD. How we manage follicular lymphoma. *Leukemia* 2014; **28:** 1388–95.

3 Salles G, Seymour JF, Offner F, et al. Rituximab maintenance for 2 years in patients with high tumour burden follicular lymphoma responding to rituximab plus chemotherapy (PRIMA): a phase 3, randomised controlled trial. *Lancet* 2011; **377:** 42–51.

4 Relander T, Johnson NA, Farinha P, Connors JM, Sehn LH, Gascoyne RD. Prognostic factors in follicular lymphoma. *J Clin Oncol* 2010; **28:** 2902–13.

5 Solal-Celigny P, Roy P, Colombat P, et al. Follicular lymphoma international prognostic index. *Blood* 2004; **104:** 1258–65.

6 Buske C, Hoster E, Dreyling M, Hasford J, Unterhalt M, Hiddemann W. The Follicular Lymphoma International Prognostic Index (FLIPI) separates high-risk from intermediate- or low-risk patients with advanced-stage follicular lymphoma treated front-line with rituximab and the combination of cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) with respect to treatment outcome. *Blood* 2006; **108:** 1504–08.

7 Kridel R, Sehn LH, Gascoyne RD. Pathogenesis of follicular lymphoma. *J Clin Invest* 2012; **122:** 3424–31.

8 Hiddemann W, Kneba M, Dreyling M, et al. Frontline therapy with rituximab added to the combination of cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) significantly improves the outcome for patients with advanced-stage follicular lymphoma compared with therapy with CHOP alone: results of a prospective randomized study of the German Low-Grade Lymphoma Study Group. *Blood* 2005; **106:** 3725–32.

9 Odejide O, Weigert O, Lane AA, et al. A targeted mutational landscape of angioimmunoblastic T-cell lymphoma. *Blood* 2014; **123:** 1293–96.

10 Pott C, Brüggemann M, Ritgen M, van der Velden VHJ, van Dongen JJM, Kneba M. MRD detection in B-cell non-Hodgkin lymphomas using Ig gene rearrangements and chromosomal translocations as targets for real-time quantitative PCR. In: Walker JM, ed. Methods in molecular biology. Clifton, NJ: Humana Press, 2012: 175–200.

11 Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014; **505:** 495–501.

12 Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2014; **499:** 214–18.

13 Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011; **39:** e118.

14 Bejar R, Stevenson K, Abdel-Wahab O, et al. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med* 2011; **364:** 2496–506.

15 Leonard RC, Hayward RL, Prescott RJ, Wang JX. The identification of discrete prognostic groups in low grade non-Hodgkin's lymphoma. The Scotland and Newcastle Lymphoma Group Therapy Working Party. *Ann Oncol* 1991; **2:** 655–62.

16 Lopez-Guillermo A, Montserrat E, Bosch F, Terol MJ, Campo E, Rozman C. Applicability of the International Index for aggressive lymphomas to patients with low-grade lymphoma. *J Clin Oncol* 1994; **12:** 1343–48.

17 Federico M, Bellei M, Marcheselli L, et al. Follicular Lymphoma International Prognostic Index 2: a new prognostic index for follicular lymphoma developed by the International Follicular Lymphoma Prognostic Factor Project. *J Clin Oncol* 2009; **27:** 4555–62.

18 Smith SM. Dissecting follicular lymphoma: high versus low risk. *Hematology Am Soc Hematol Educ Program* 2013; **2013:** 561–67.

19 Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56:** 779–88.

20 Yap DB, Chu J, Berg T, et al. Somatic mutations at EZH2 Y641 act dominantly through a mechanism of selectively altered PRC2 catalytic activity, to increase H3K27 trimethylation. *Blood* 2011; **117:** 2451–59.

21 McCabe MT, Graves AP, Ganji G, et al. Mutation of A677 in histone methyltransferase EZH2 in human B-cell lymphoma promotes hypertrimethylation of histone H3 on lysine 27 (H3K27). *Proc Natl Acad Sci USA* 2012; **109:** 2989–94.

22 Lenz G, Davis RE, Ngo VN, et al. Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* 2008; **319:** 1676–79.

23 Hunter ZR, Xu L, Yang G, et al. The genomic landscape of Waldenstrom macroglobulinemia is characterized by highly recurring MYD88 and WHIM-like CXCR4 mutations, and small somatic deletions associated with B-cell lymphomagenesis. *Blood* 2014; **123:** 1637–46.

24 Trinh DL, Scott DW, Morin RD, et al. Analysis of FOXO1 mutations in diffuse large B-cell lymphoma. *Blood* 2013; **121:** 3666–74.

25 Yildiz M, Li H, Bernard D, et al. Activating STAT6 mutations in follicular lymphoma. *Blood* 2015; **125:** 668–79.

26 Ying CY, Dominguez-Sola D, Fabi M, et al. MEF2B mutations lead to deregulated expression of the oncogene BCL6 in diffuse large B cell lymphoma. *Nat Immunol* 2013; **14:** 1084–92.

27 Khodabakhshi AH, Morin RD, Fejes AP, et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* 2012; **3:** 1308–19.

28 Leich E, Salaverria I, Bea S, et al. Follicular lymphomas with and without translocation t(14;18) differ in gene expression profiles and genetic alterations. *Blood* 2009; **114:** 826–34.

29 Bödör C, Grossmann V, Popov N, et al. EZH2 mutations are frequent and represent an early event in follicular lymphoma. *Blood* 2013; **122:** 3165–68.

30 O'Shea D, O'Riain C, Taylor C, et al. The presence of TP53 mutation at diagnosis of follicular lymphoma identifies a high-risk group of patients with shortened time to disease progression and poorer overall survival. *Blood* 2008; **112:** 3126–29.

31 Cheung KJ, Johnson NA, Affleck JG, et al. Acquired TNFRSF14 mutations in follicular lymphoma are associated with worse prognosis. *Cancer Res* 2010; **70:** 9166–74.

32 Launay E, Pangault C, Bertrand P, et al. High rate of TNFRSF14 gene alterations related to 1p36 region in de novo follicular lymphoma and impact on prognosis. *Leukemia* 2012; **26:** 559–62.

33 Correia C, Schneider PA, Dai H, et al. BCL2 mutations are associated with increased risk of transformation and shortened survival in follicular lymphoma. *Blood* 2015; **125:** 658–67.

34 McCabe MT, Ott HM, Ganji G, et al. EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* 2013; **492:** 108–12.

35 Béguelin W, Popovic R, Teater M, et al. EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell* 2013; **23:** 677–92.

3.2 Publication 2 - *Blood*, 2016

Jurinovic V, Kridel R, Staiger AM, et al. Clinicogenetic risk models predict early progression of follicular lymphoma after first-line immunochemotherapy. *Blood*, 128(8):1112-1120, 2016.

Regular Article

blood

LYMPHOID NEOPLASIA

Clinicogenetic risk models predict early progression of follicular lymphoma after first-line immunochemotherapy

Vindi Jurinovic,[1,2] Robert Kridel,[3] Annette M. Staiger,[4,5] Monika Szczepanowski,[6] Heike Horn,[4,5] Martin H. Dreyling,[1] Andreas Rosenwald,[7] German Ott,[4] Wolfram Klapper,[6] Andrew D. Zelenetz,[8] Paul M. Barr,[9] Jonathan W. Friedberg,[9] Stephen Ansell,[10] Laurie H. Sehn,[3] Joseph M. Connors,[3] Randy D. Gascoyne,[3] Wolfgang Hiddemann,[1,11] Michael Unterhalt,[1] David M. Weinstock,[12] and Oliver Weigert[1,11]

[1]Medical Department III, University Hospital of the Ludwig Maximilians University Munich, Munich, Germany; [2]Institute for Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University Munich, Munich, Germany; [3]Centre for Lymphoid Cancer, British Columbia Cancer Agency, Vancouver, BC, Canada; [4]Department of Clinical Pathology, Robert-Bosch-Krankenhaus, Stuttgart, Germany; [5]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Tübingen, Germany; [6]Hematopathology Section, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany; [7]Institute of Pathology, University of Würzburg, and Comprehensive Cancer Center Mainfranken, Würzburg, Germany; [8]Department of Medicine, Memorial Sloan Kettering Cancer Center, and Weill Medical College of Cornell University, New York, NY; [9]Wilmot Cancer Institute, University of Rochester, Rochester NY; [10]Division of Hematology, Mayo Clinic, Rochester, MN; [11]German Cancer Consortium and German Cancer Research Center, Heidelberg, Germany; and [12]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA

Key Points

- The posttreatment end point progression of FL within 24 months (POD24) is strongly associated with OS.
- A pretreatment clinicogenetic risk model (m7-FLIPI) predicts POD24 and OS and identifies the smallest subgroup with highest unmet need.

Follicular lymphoma (FL) is a clinically and molecularly heterogeneous disease. Posttreatment surrogate end points, such as progression of disease within 24 months (POD24) are promising predictors for overall survival (OS) but are of limited clinical value, primarily because they cannot guide up-front treatment decisions. We used the clinical and molecular data from 2 independent cohorts of symptomatic patients in need of first-line immunochemotherapy (151 patients from a German Low-Grade Lymphoma Study Group [GLSG] trial and 107 patients from a population-based registry of the British Columbia Cancer Agency [BCCA]) to validate the predictive utility of POD24, and to evaluate the ability of pretreatment risk models to predict early treatment failure. POD24 occurred in 17% and 23% of evaluable GLSG and BCCA patients, with 5-year OS rates of 41% (vs 91% for those without POD24, $P < .0001$) and 26% (vs 86%, $P < .0001$), respectively. The m7–FL International Prognostic Index (m7-FLIPI), a prospective clinicogenetic risk model for failure-free survival, had the highest accuracy to predict POD24 (76% and 77%, respectively) with an odds ratio of 5.82 in GLSG ($P = .00031$) and 4.76 in BCCA patients ($P = .0052$). A clinicogenetic risk model specifically designed to predict POD24, the POD24-PI, had the highest sensitivity to predict POD24, but at the expense of a lower specificity. In conclusion, the m7-FLIPI prospectively identifies the smallest subgroup of patients (28% and 22%, respectively) at highest risk of early failure of first-line immunochemotherapy and death, including patients not fulfilling the POD24 criteria, and should be evaluated in prospective trials of precision medicine approaches in FL. (*Blood*. 2016; 128(8):1112-1120)

Introduction

Follicular lymphoma (FL) is among the most common malignant lymphomas worldwide and remains incurable for most patients.[1] FL is a highly heterogeneous disease,[2] with a subgroup of patients experiencing remarkably poor outcome. Several recent studies have suggested that posttreatment surrogate end points are powerful predictors for overall survival (OS).[3,4] For example, 19% to 26% of patients receiving first-line immunochemotherapy with rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) experienced progression of disease within 24 months ("early progression of disease,"[4] herein referred to as POD24) and had a 5-year OS of only 34% to 50% compared with a 5-year OS of 90% to 94% for patients without POD24.[4] Independent validation of these results is needed, also in the context of different treatment regimens. Furthermore, the length of first remission was calculated differently across studies, either from date of diagnosis[4] (for database reasons) or after treatment.[3]

Although conceptually similar results are emerging for event-free survival at 12 and 24 months (EFS12 and EFS24)[5,6] and complete response rate at 30 months (CR30),[7] retrospective evaluation of treatment outcome is of limited clinical utility, because it cannot be used to guide up-front treatment decision. Furthermore, the molecular determinants of poor patient outcome remain to be defined. To develop precision medicine treatment strategies, it is essential to establish pretreatment strategies for risk assessment that include clinically relevant biomarkers.

Submitted May 17, 2016; accepted July 5, 2016. Prepublished online as *Blood* First Edition paper, July 14, 2016; DOI 10.1182/blood-2016-05-717355.

The online version of this article contains a data supplement.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

We have previously shown that a clinicogenetic risk model called the m7–FL International Prognostic Index (m7-FLIPI), which includes the mutation status of 7 genes (*EZH2*, *ARID1A*, *MEF2B*, *EP300, FOXO1, CREBBP*, and *CARD11*), the FLIPI, and the Eastern Cooperative Oncology Group (ECOG) performance status at the time of treatment initiation improves risk stratification for failure-free survival (FFS) in patients with FL receiving first-line immunochemotherapy.[8] An online tool for calculating the m7-FLIPI is available at: http://www.glsg.de/m7-flipi/.

In this study, we aimed to independently validate the predictive utility of posttreatment evaluation by POD24 in 2 independent cohorts of patients who received different immunochemotherapy regimens as first-line treatment. Furthermore, we evaluated and compared the ability of pretreatment risk models, including the m7-FLIPI, to predict POD24, and explored additional pretreatment risk models specifically designed to predict POD24.

Methods

We fully reanalyzed the clinical and molecular data from 2 independent cohorts of patients with symptomatic, advanced stage, or bulky FL considered ineligible for curative radiotherapy. All patients had an available biopsy specimen obtained within 12 months before the initiation of first-line therapy that was previously sequenced to determine the mutational status of 74 genes.[8]

Briefly, the GLSG cohort consisted of 151 patients who needed treatment as defined by the presence of B-symptoms, bulky disease (mediastinal lymphomas >7.5 cm or other lymphomas >5 cm), impairment of normal hematopoiesis (hemoglobin level <100 g/L, granulocyte count $<1.5 \times 10^9$/L, or thrombocyte count $<100 \times 10^9$/L), compression of internal organs, or disease progression (>50% increase of lymphoma manifestations within <6 months). All patients received R-CHOP and interferon-α (IFN-α) maintenance as part of the randomized GLSG2000 trial of the German Low-Grade Lymphoma Study Group (GLSG).[9] Median age of GLSG patients was 57 years (range 27-77); 77 (51%) had high-risk FLIPI. With a median follow-up of 7.7 years, 5-year FFS and OS rates were 66% and 83%, respectively.[8]

The BCCA cohort consisted of 107 patients from a population-based registry of the British Columbia Cancer Agency (BCCA) who received rituximab, cyclophosphamide, vincristine, and prednisone (R-CVP), followed by R-maintenance by intention to treat in 93 patients (87%). Median age of BCCA patients was 62 years (range 37-83); 53 (50%) had high-risk FLIPI. With a median follow-up of 6.7 years, 5-year FFS and OS rates were 58% and 74%, respectively.[8]

Progression of disease within 24 months was defined as progression or relapse of the disease within the first 24 months after diagnosis (original definition)[4] or after first-line treatment initiation (modified definition). Patients were not evaluable for POD24 if they were censored (eg, lost to follow-up) or died within 24 months without POD.

Failure-free survival was defined as time from treatment initiation until less than a partial remission (PR) at the end of induction, relapse, progression, or death from any cause. Overall survival was calculated from risk-defining event for POD24 (ie, survival from time of POD for the POD24 cohort, or from 2 years after initial treatment of patients without POD24),[4] and from treatment initiation for all other survival analyses.

Clinical and molecular data from the GLSG cohort were used to calculate a risk model that specifically predicts POD24 (POD24 Prognostic Index [POD24-PI]) by applying a previously described statistical approach.[8] Briefly, the mutation status of genes that were mutated in >5 patients and the clinical risk factors FLIPI >2 (ie, high-risk FLIPI) and poor performance status (ECOG-PS >1) were used for multivariable L1-penalized logistic regression. Two different risk models were calculated. In the first model, the coefficients for high-risk FLIPI and ECOG-PS >1 were not penalized, forcing these variables into the model. In the second model, all coefficients were penalized. Internal validation by the bootstrap procedure was used to select the best model. The final risk score was calculated as the sum of clinical and molecular predictors weighted by their individual Lasso coefficients. We determined the optimal cutoff value to maximize the Wald statistic, and dichotomized patients into high-risk and low-risk subgroups. The BCCA cohort was used as an independent validation cohort.

Logistic regression analyses were performed to assess whether risk models were predictive of POD24, and Cox regression analysis was used for FFS and OS. All calculations were carried out with the statistical software R (version 3.1.2). The accuracy of pretreatment risk models to predict POD24 was calculated as the number of correctly classified patients ([number of true positives + number of true negatives] ÷ [number of all evaluable patients]). The R-package *penalized* (version 0.9-45) was used for penalized logistic regression, and the *survival* package (version 2.37-7) for survival analyses.

This study was covered by approvals of the Ludwig-Maximilians-University Munich Institutional Review Board (#056-00) and the University of British Columbia–BCCA Research Ethics Board (#H13- 01765).

Results

Validation of POD24 to identify high-risk patients

We first aimed to assess the prognostic impact of POD24 on OS in 2 independent cohorts of patients with FL receiving first-line immunochemotherapy. Nineteen (13%) and 5 patients (5%) from the entire GLSG and BCCA cohorts were not evaluable for analysis of POD24 because they were censored or died within 24 months without prior POD (Figure 1A). POD24, originally defined as relapse or progression of FL within 24 months of diagnosis,[4] occurred in 15% (20/132) and 18% (18/102) of evaluable patients from the GLSG and the BCCA cohorts (Figure 1A). When calculated from time of first-line treatment initiation to overcome the lead-time bias (ie, the time between diagnosis and symptomatic disease requiring treatment), the size of POD24 subgroups increased to 17% (23/132) and 23% (23/102), respectively (Figure 1A; Table 1). Only 1 of the 8 reclassified patients was still alive at 7.6 years, and the median OS of this subgroup was only 3.1 years (range 1.4-9.5; $P < .0001$ compared with all other patients). The number of reclassified patients is small (the time between diagnosis and treatment was <1 year by inclusion criteria), but the poor outcome of patients with POD within 24 months of treatment, but not from diagnosis, suggests that these patients should also be considered early progressors. Thus, the modified definition of POD24 was used for the remainder of the study.

Differences in OS, calculated from risk-defining event (ie, survival from time of POD for early progressors, or from 24 months after initial treatment for non–early progressors) were highly significant between patients with and without POD24 (irrespective of whether original or modified definitions were used; Figure 1B). Six and 4 patients with POD24 from the GSLG and BCCA cohorts were still alive at 5 years, for a 5-year OS of 41% vs 91% (hazard ratio [HR] 9.72, 95% confidence interval [CI] [4.51; 20.96], $P < .0001$) and 26% vs 86% (HR 11.93, 95% CI [5.31; 26.76], $P < .0001$), respectively (Figure 1B). This confirms that retrospective evaluation of treatment response at 24 months is strongly associated with OS in patients receiving first-line immunochemotherapy.

The clinical characteristics of the POD24 and non-POD24 subgroups are summarized in Table 1. The POD24 subgroups were enriched for high-risk FLIPI (78% vs 44%, $P = .0059$, and 70% vs 42%, $P = .035$, for GLSG and BCCA patients, respectively; Table 1; Figure 2A). Another 44% and 42% of patients without POD24 had high-risk FLIPI, respectively (Figure 2A). However, patients with high-risk FLIPI and no POD24 did not have inferior FFS compared with those with low-risk FLIPI and no POD24 (Figure 3A). This

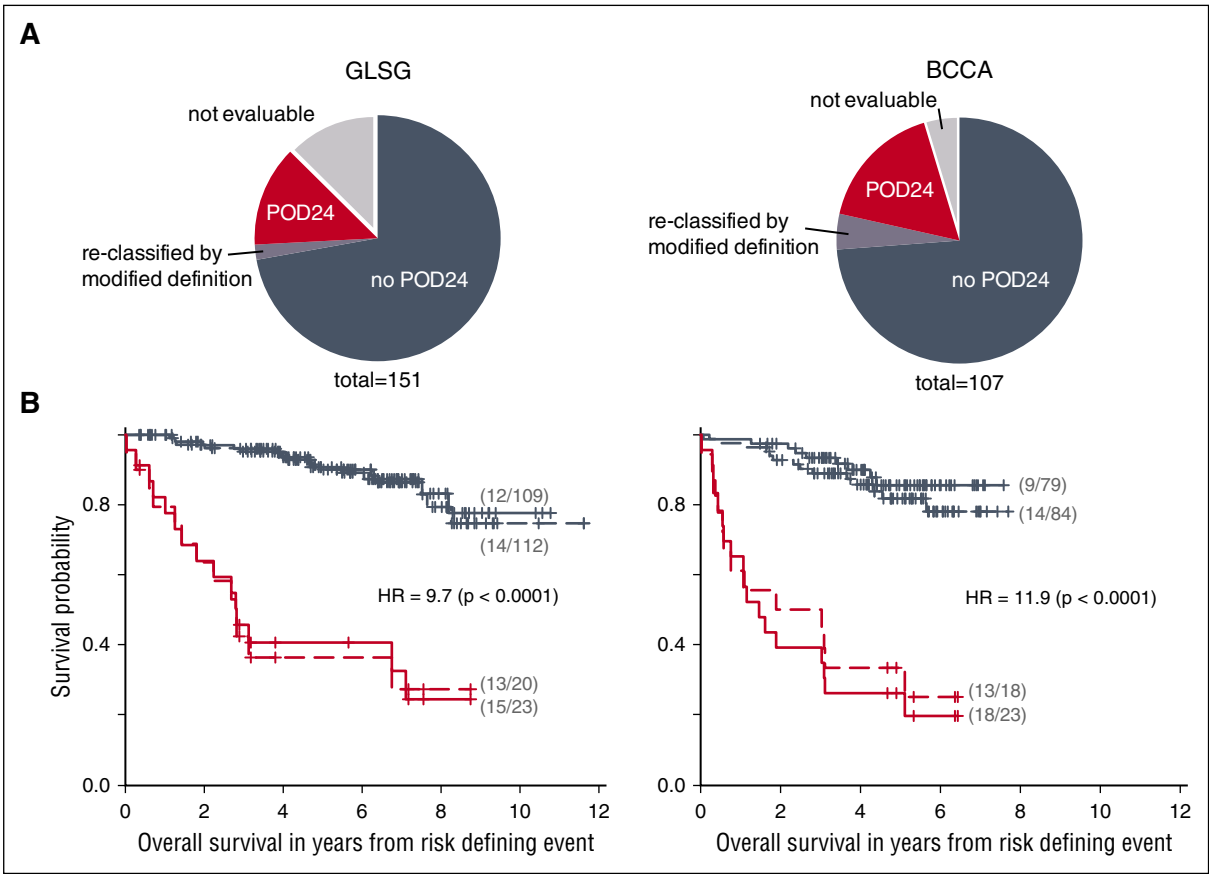

Figure 1. Progression of disease within 24 months (POD24) is an accurate predictor of poor overall survival (OS). (A) Distribution of patients from the GLSG and BCCA cohorts according to the POD24 classifier. (B) Kaplan Meier curves for OS from risk-defining events for patients with or without POD24 of diagnosis (dashed lines) or 24 months of treatment initiation (solid lines) from the GLSG (left) and BCCA cohorts (right). Displayed statistics refer to POD24 status calculated from time of treatment initiation (ie, modified definition).

suggests that the FLIPI, which uses only clinical factors and classified 51% of GLSG and 50% of BCCA patients as high-risk, overestimates the number of patients with poor outcome.

The m7-FLIPI is predictive of POD24

We have previously shown that by integrating the mutation status of 7 genes with clinical risk factors, the clinicogenetic risk model m7-FLIPI results in reclassification of approximately one half of high-risk FLIPI patients into the low-risk group.[8] Now, we assessed the performance of the m7-FLIPI to prospectively distinguish patients with and without POD24.

Unlike the POD24 classifier, all patients were evaluable for the m7-FLIPI (Table 2). Forty-three GLSG (28%) and 24 BCCA patients (22%) were classified as high risk by the m7-FLIPI, with a 5-year OS from treatment initiation of 65% vs 90% (HR 3.4, $P < .0001$) and 42% vs 84% (HR 4.9, $P < .0001$), respectively (Table 2). High-risk m7-FLIPI patients were significantly more likely to develop POD24

Table 1. Patient and disease characteristics according to POD24 status

	GLSG			BCCA		
	POD24	No POD24	*P*	POD24	no POD24	*P*
No. of evaluable patients	23	109		23	79	
First-line treatment		R-CHOP (151/151, 100%)			R-CVP (107/107, 100%)	
Maintenance treatment by ITT		IFN (151/151, 100%)			Rituximab (93/107, 87%)	
Median follow-up in years	8.4	8.2		7.1	6.7	
Age (y), median (range)	61 (27-74)	56 (29-77)	.195	62 (43-83)	61 (37-83)	.398
Male gender	11/23 (48%)	55/109 (50%)	>.99	15/23 (65%)	42/79 (53%)	.618
High-risk FLIPI	18/23 (78%)	48/109 (44%)	.0059	16/23 (70%)	33/79 (42%)	.035
Age >60 y	12/23 (52%)	39/109 (36%)	.218	12/23 (52%)	42/79 (53%)	>.99
No. of nodal sites >4	19/23 (83%)	71/109 (65%)	.165	20/23 (87%)	55/79 (70%)	.164
LDH elevated	11/23 (48%)	31/109 (28%)	.117	6/21 (29%)	15/77 (19%)	.548
Hb <120 g/L	10/23 (43%)	17/109 (16%)	.0064	4/21 (19%)	7/79 (9%)	.350
ECOG-PS ≥2	0/23 (0%)	5/109 (5%)	.655	6/23 (26%)	9/79 (11%)	.157

ECOG-PS, Eastern Cooperative Oncology Group Performance Status; Hb, hemoglobin; IFN, interferon-α; ITT, intention-to-treat; LDH, lactate dehydrogenase; POD24, progression of disease within 24 months; R-CHOP, rituximab, cyclophosphamide, doxorubicin, vincristine, prednisone; R-CVP, rituximab, cyclophosphamide, vincristine, prednisone.

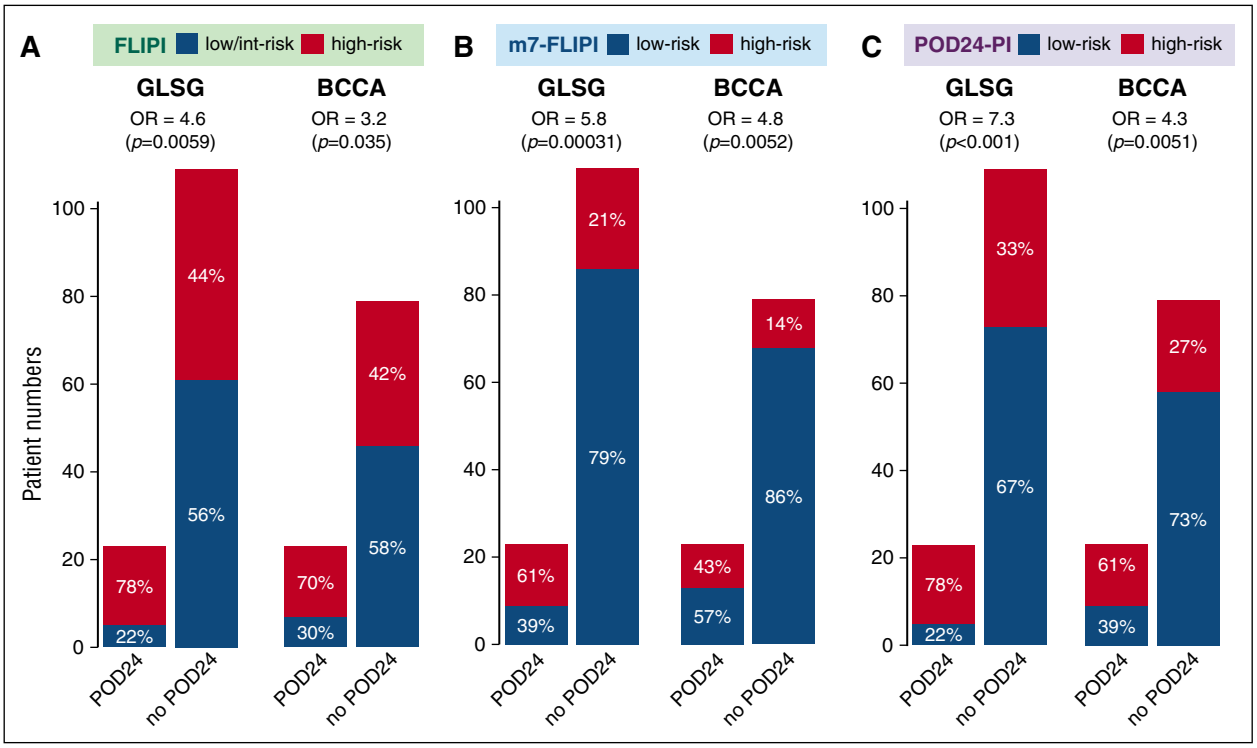

Figure 2. Accuracy of 3 pretreatment risk models to predict POD24 status. (A) Follicular Lymphoma International Prognostic Index (FLIPI), (B) m7-FLIPI, and (C) POD24 Prognostic Index (POD24-PI).

with an odds ratio (OR) of 5.82 (95% CI [2.27; 15.63]; $P = .00031$) and 4.76 (95% CI [1.68; 13.72]; $P = .0052$) in GLSG and BCCA patients (Figure 2B; supplemental Figure 1A). Compared with the FLIPI, the specificity of the m7-FLIPI to identify POD24 (ie, the true negative rate) increased from 56% to 79%, and 58% to 86%, respectively (Figure 2B). However, 21% of GLSG and 14% of BCCA patients who did not experience POD24 were still assigned into the high-risk m7-FLIPI subgroup (Figure 2B). To determine whether these cases have an inferior prognosis even though they do not progress within 24 months, we analyzed the impact of high-risk m7-FLIPI in patients without POD24. In both cohorts, high-risk m7-FLIPI was still associated with a shorter FFS (Figure 3B) and OS (supplemental Figure 2B) among patients who did not have POD24. Thus, the accuracy of the FLIPI to predict POD24 is substantially improved by adding the ECOG-PS and the mutation status of 7 genes (Table 2). Furthermore, the m7-FLIPI is also predictive for treatment outcome in patients not fulfilling the criteria of POD24.

A clinicogenetic risk classifier specifically designed to predict POD24

Despite the superior performance of the m7-FLIPI, 6% of GLSG patients (9/151) and 12% of BCCA patients (13/107) were classified as low-risk m7-FLIPI but developed progression of FL within 24 months of treatment (4 and 6 of whom were high-risk FLIPI), for an overall sensitivity of 61% and 43%, respectively, at predicting POD24 (Figure 2B). We aimed to improve that by using the clinical and molecular data from the GLSG cohort to calculate another risk model that specifically predicts POD24. Internal validation showed superiority of the model in which all clinical and molecular coefficients were penalized (bootstrap-corrected coefficient of 0.95 vs 0.23 for the model in which the coefficients for high-risk FLIPI and ECOG-PS >1 were not penalized). We termed this risk model the POD24 Prognostic Index (POD24-PI). The risk score, calculated as the sum of predictor values weighted by Lasso coefficients, contained 4 factors that were all within the m7-FLIPI: high-risk FLIPI ($\beta_{Lasso}=1.0$), and nonsilent mutations in *EP300* ($\beta_{Lasso}=0.58$), *FOXO1* ($\beta_{Lasso}=0.14$), and *EZH2* ($\beta_{Lasso}=-0.42$) (Figure 4A). The optimal cutoff value to stratify patients into high- and low-risk subgroups was determined to be 0.71 (Figure 4A). The BCCA cohort was used to independently validate the results.

Compared with the m7-FLIPI, a higher fraction of patients was classified into the high-risk subgroup by the POD24-PI (Figures 2C and 5), specifically 42% (63/151) and 36% (39/107) of GLSG and BCCA patients, respectively (Table 2). As intended, the POD24-PI had a higher sensitivity to predict POD24 compared with the m7-FLIPI (78% vs 61%, and 61% vs 43% in the GLSG and BCCA cohorts, respectively; Figure 2C), albeit at the cost of a lower specificity and accuracy (Table 2; supplemental Figure 1B). Overall, high-risk POD24-PI was associated with significantly shorter FFS and OS (Figure 4; Table 2): the 5-year FFS rates were 50% vs 77% (HR = 3.06, $P < .0001$) and 36% vs 72% (HR = 3.01, $P < .0001$), and the 5-year OS rates were 71% vs 91% (HR = 3.55, $P = .00026$) and 48% vs 89% (HR = 5.35, $P < .0001$) in the GLSG and BCCA cohorts, respectively (Figure 4). In patients without POD24, high-risk POD24-PI was still associated with a shorter FFS and OS, but less discriminative compared with the m7-FLIPI (ie, the POD24-PI had lower HRs and inferior *P* values compared with the m7-FLIPI (Figure 3C; supplemental Figure 2C).

Table 2 summarizes the specific features of the 2 clinicogenetic risk scores, in context with the FLIPI and the POD24 classifier. Although the m7-FLIPI had the highest accuracy and POD24-PI the highest sensitivity to predict POD24, 22% (5/23) and 30% (7/23) of patients with POD24 from the GLSG and BCCA cohorts were still not correctly identified as high risk by any of the pretreatment risk models (Figure 5). Because mutations in *TP53* are not included in any of the clinicogenetic risk models but are known to be associated

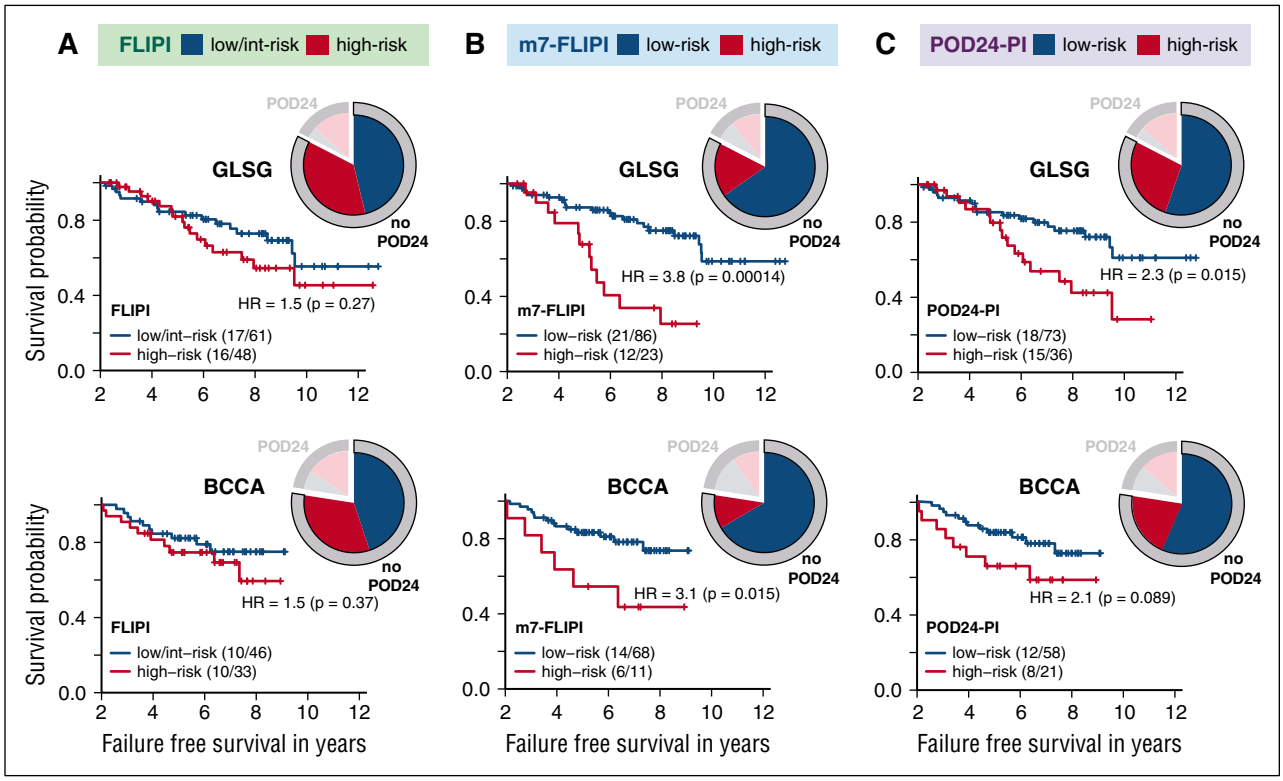

Figure 3. Risk stratification for failure-free survival (FFS) in patients without POD24 according to 3 pretreatment risk models. Kaplan-Meier curves for FFS beyond 2 years after treatment initiation for patients without POD24 according to (A) the FLIPI, (B) the m7-FLIPI, and (C) the POD24-PI. Numbers in parentheses indicate patients with event/number of patients per subgroup. Pie charts illustrate distribution of risk status of the respective risk classifier and POD24 status.

with inferior OS,[8,10] we compared *TP53* mutation frequency in patients with or without POD24. In both cohorts, *TP53* mutations were in fact enriched in the POD24 subgroup (13% [3/23] vs 3% [3/109] in GLSG patients [$P = .11$], and 13% [3/23] vs 4% [3/79] in BCCA patients [$P = .25$]), but failed to reach statistical significance (supplemental Table 1; supplemental Figure 3).

Discussion

Currently applied immunochemotherapy regimens result in long-lasting remissions and excellent OS in a majority (~80%) of patients with FL requiring systemic treatment. However, our study confirms that a subset of patients (~20%) experience short remissions and markedly inferior outcome with a median OS of <5 years. Clearly, strategies to guide risk-adapted treatment approaches in FL are needed to avoid overtreatment of low-risk patients, and to prioritize alternative over standard treatment regimens in high-risk patients. Also, clinical trials focusing on high-risk patients are likely to identify higher activity regimens at a much faster rate if study results were not mitigated by patients with highly indolent clinical courses in unselected study cohorts.

Retrospective evaluation of treatment response at 24 months after first-line immunochemotherapy currently represents the strongest predictor of OS, although a subset of patients with POD24 are still alive at >5 years (26% and 41% in our series, up to 50% in a previous series[4]). By its definition, POD24 is not confounded by subsequent therapies (as is OS), or by deaths without prior POD as a result of comorbidity or treatment-related mortality (as is progression-free survival or event-free survival),[11] and thus very closely reflects either the aggressiveness of the disease and/or treatment-specific resistance. As such, POD24 will be highly useful to select cases for in-depth molecular characterization to identify the tumor-biological determinants of poor patient outcome.

POD24 will immediately be useful in clinical practice to select high-risk patients for experimental salvage treatments. One such example is the S1608 trial conducted through the National Cancer Institute's National Clinical Trial Network, which will specifically enroll patients with POD24 after first-line immunochemotherapy. However, as a posttreatment surrogate marker, POD24 cannot guide first-line treatment including consolidation/maintenance regimens in first remission, and by definition is unable to assess patients who die within 24 months without prior documented POD or to identify high-risk patients who do not fail first-line treatment within 24 months.

We propose that comprehensive risk models that integrate established clinical risk factors with disease-specific biomarkers to predict biology-relevant end points are useful in up-front identification of high-risk patients. The previously described m7-FLIPI is the most stringent pretreatment risk model currently available and identifies the smallest subgroup of patients (~25%) at highest risk of early failure of first-line immunochemotherapy and death. The m7-FLIPI has the highest accuracy and PPV for POD24 among all pretreatment risk models. Also, high-risk m7-FLIPI is associated with inferior outcome in patients who do not fail treatment within 24 months, a subset currently missed by the POD24 classifier. As such, high-risk m7-FLIPI prospectively defines the subgroup of patients with the highest clinical need in FL before initiation of first-line treatment, and supports clinical trials with alternative up-front regimens with highest antitumor activity, potentially accepting higher toxicity profiles as deemed acceptable for the majority of patients with low-risk disease. Furthermore, among all pretreatment risk scores, the m7-FLIPI has the highest specificity for POD24 (ie, it identifies the

Table 2. Features and test metrics of tested risk classifiers for previously untreated patients with symptomatic, advanced stage FL from 2 independent cohorts

	POD24*	FLIPI	m7-FLIPI	POD24-PI
Type	Posttreatment	Pretreatment	Pretreatment	Pretreatment
Primary endpoint	OS	OS	FFS	POD24
Validated on independent cohort	Yes	Yes	Yes	Yes
Clinical predictors	POD24	Age >60 y, no. of nodal sites >3, elevated serum LDH, hemoglobin <12 g/L, Ann Arbor stage III/IV	High-risk FLIPI, ECOG performance status >2	High-risk FLIPI
Molecular predictors	None	None	Nonsilent mutations in 7 genes (*ARID1A, CARD11, CREBBP, EP300, EZH2, FOXO1, MEF2B*) at VAFs ≥10%	Nonsilent mutations in 3 genes (*EP300, EZH2, FOXO1*) at VAFs ≥10%
Calculation of risk score	Single variable	Cumulative sum of predictor values, all predictors have weight = 1	Cumulative sum of predictor values, predictors have individual weights†	Cumulative sum of predictor values, predictors have individual weights (Figure 4A)
Number of risk groups	2 (no POD24 = low-risk, POD24 = high-risk)	3 (0-1 = low-risk, 2 = interm-risk, 3-5 = high-risk)	2 (<0.8 = low-risk, >0.8 = high-risk)	2 (<0.71 = low-risk, >0.71 = high-risk)
High-risk cases among patients with symptomatic, advanced stage FL (%)	GLSG: 23/132 (17) BCCA: 23/102 (22)	GLSG: 77/151 (51) BCCA: 53/107 (50)	GLSG: 43/151 (28) BCCA: 24/107 (22)	GLSG: 63/151 (42) BCCA: 39/107 (36)
Predictive for POD24	n/a	Yes	Yes	Yes
GLSG: evaluable patients 132/151		OR = 4.6 (P = .0059) Sens 78%, Spec 56%, AUC 0.67 Accuracy 60%, PPV 27%, NPV 92%	OR = 5.8 (P = .00031) Sens 61%, Spec 79%, AUC 0.70 Accuracy 76%, PPV 38%, NPV 91%	OR = 7.3 (P = .00016) Sens 78%, Spec 67%, AUC 0.73 Accuracy 71%, PPV 33%, NPV 94%
BCCA: evaluable patients 107/107		OR = 3.2 (P = .035) Sens 70%, Spec 58%, AUC 0.64 Accuracy 61%, PPV 33%, NPV 87%	OR = 4.8 (P = .0052) Sens 43%, Spec 86%, AUC 0.65 Accuracy 77%, PPV 48%, NPV 84%	OR = 4.3 (P = .0051) Sens 61%, Spec 73%, AUC 0.67 Accuracy 69%, PPV 40%, NPV 87%
Predictive for FFS	n/a	Yes, but only as a binary classifier (low-/interm-risk vs high-risk)	Yes	Yes
GLSG: evaluable patients 151/151		5-y FFS 57% vs 76% HR = 2.11 (P = .0034)	5-y FFS 38% vs 77% HR = 4.14 ($P = 6.313 \cdot 10^{-9}$)	5-y FFS 50% vs 77% HR = 3.06 ($P = 5.989 \cdot 10^{-6}$)
BCCA: evaluable patients 107/107		5-y FFS 47% vs 70% HR = 2.18 (P = .0075)	5-y FFS 25% vs 68% HR = 3.58 ($P = 4.924 \cdot 10^{-6}$)	5-y FFS 36% vs 72% HR = 3.01 ($P = 7.178 \cdot 10^{-5}$)
Predictive for OS	Yes‡	Yes	Yes	Yes
GLSG: evaluable patients 151/151	5-y OS 40% vs 96% HR = 10.91 ($P = 5.262 \cdot 10^{-14}$)	5-y OS 75% vs 91% HR = 2.59 (P = .0083), C-Index 0.75	5-y OS 65% vs 90% HR = 3.38 (P = .00031), C-Index 0.78	5-y OS 71% vs 91% HR = 3.55 (P = .00026), C-Index 0.79
BCCA: evaluable patients 107/107	5-y OS 26% vs 93% HR = 13.602 ($P = 6.661 \cdot 10^{-16}$)	5-y OS 60% vs 89% HR = 3.90 (P = .00034), C-Index 0.81	5-y OS 42% vs 84% HR = 4.89 ($P = 7.661 \cdot 10^{-7}$), C-Index 0.84	5-y OS 48% vs 89% HR = 5.35 ($P = 9.996 \cdot 10^{-7}$), C-Index 0.86
Comments	Risk classification not possible for patients who died/were censored within 24 mo of first-line treatment. Only predictive for OS. First described in 2015.	Most widely used and best validated pretreatment classifier. Contains only clinical variables not necessarily directly reflecting disease biology. First described in 2004, not widely used to guide treatment decisions.	Highest accuracy and specificity to predict POD24. Most discriminative classifier for patients without POD24 into high- and low-risk groups. Sequencing of 7 genes required. First described in 2015, requires further validation.	Highest sensitivity to predict POD24. Sequencing of 3 genes required. Requires further validation.

AUC, area under the curve; CR30, complete response rate at 30 months; FLIPI, Follicular Lymphoma International Prognostic Index; HR, hazard ratio; NPV, negative predictive value; OR, odds ratio; POD, progression of disease; POD24-PI, early progression prognostic index; PPV, positive predictive value; Sens, sensitivity; Spec, specificity.

*The modified definition of POD24 was used for this analysis (see text).

†http://www.glsg/m7-FLIPI.

‡For this analysis only, OS for POD24 was calculated from time of treatment initiation for better comparison (nota bene, numbers differ from the text, wherein OS was calculated from time of risk-defining event).

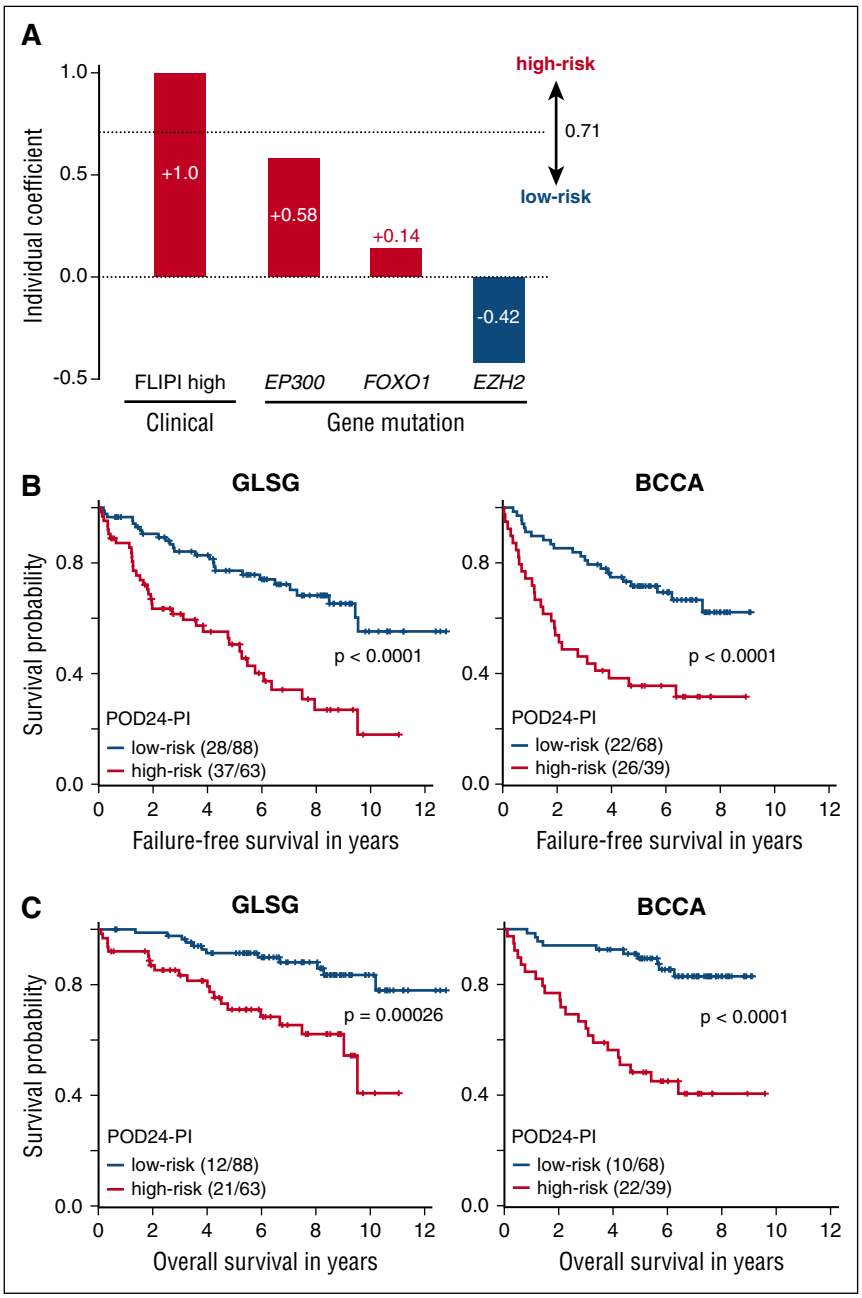

Figure 4. The POD24 Prognostic Index (POD24-PI). (A) The POD24-PI is calculated as the sum of individual clinical and gene mutation predictor values weighted by their individual coefficients. (B) Kaplan-Meier curves for failure-free survival (FFS), and (C) overall survival (OS) for patients from the GLSG and BCCA cohorts according to POD24-PI status. Numbers in parentheses indicate patients with event/number of patients per subgroup.

highest percentage of non–early progressors correctly as low-risk). This indicates that the m7-FLIPI might also be useful in up-front identification of low-risk patients with excellent outcome with currently applied immunochemotherapy regimens, and a subset might actually qualify for treatment de-escalation strategies.

The POD24-PI, specifically designed to improve the sensitivity to predict POD24, classified more patients into the high-risk subgroup (~40%), which was less enriched for poor outcome compared with high-risk m7-FLIPI. Despite its inferior performance by most test metrics, the POD24-PI may still be considered a valuable predictor in certain clinical situations; eg, when testing very-well-tolerated regimens (eg, post-remission vaccines) investigators may want to minimize the risk of excluding high-risk patients while accepting some that have been falsely identified as such. Furthermore, the fact that the POD24-PI contains the 4 highest weighted components of the m7-FLIPI likely explains the performance of the latter to predict POD24, and provides clues about how the biology of high-risk tumors may be different from others. Of note, a subset of patients with POD24 was not distinguishable by any of the 2 clinicogenetic risk models, suggesting that further improvements and probably integration of additional biomarkers are needed to capture these cases.

Based on the results from the PRIMA trial,[12] many patients now receive maintenance treatment with rituximab after first-line immunochemotherapy. Interestingly and similar to previous studies,[4] the percentage of patients progressing within 24 months was in the 20% range in both of our cohorts, despite IFN maintenance in GLSG patients and rituximab maintenance for the majority of BCCA patients, implying no major impact of these approaches on POD24. Thus, substantial improvement of treatment results is most likely to be expected from innovative, risk-adapted first-line regimens, eventually combined with minimal residual disease–guided consolidation/maintenance strategies.[13-17]

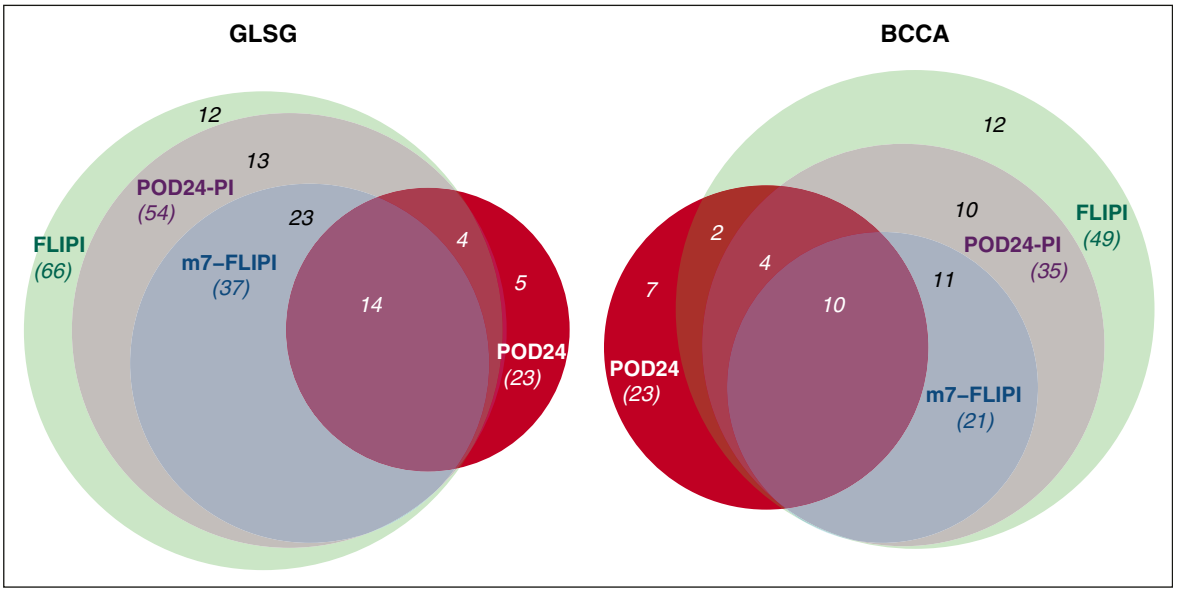

Figure 5. Identification of high-risk patients by 3 pretreatment risk models and POD24 status. Venn diagrams visualize the overlap of high-risk patients as identified by the FLIPI, m7-FLIPI, POD24-PI, and POD24.

In this study, we analyzed stringently selected patients with advanced stage or bulky disease in need of systemic treatment from both a prospective clinical trial (the GLSG cohort), which might not necessarily reflect routine clinical practice,[18] and a population-based registry (the BCCA cohort), a retrospective cohort that might be more prone to confounding and bias, but also more closely reflects real-life patients. Remarkably, analyzing these 2 different cohorts yielded highly consistent results. As such, the m7-FLIPI establishes solid grounds for up-front patient stratification by actual risk; however, several challenges still remain to be addressed before it can be applied in clinical trials and practice. Standardization of molecular technologies and analysis pipelines will be needed to ensure widely reproducible results. The m7-FLIPI will have to be validated and compared with other posttreatment surrogate markers (eg, EFS12, EFS24, and CR30)[5-7] and pretreatment risk models (eg, the FLIPI-2)[19] in additional and larger cohorts with longer follow-up, and evaluated in the context of specific treatments, such as the now widely used bendamustine plus rituximab regimen.[20,21] Integrating gene mutations into risk assessment for molecular-targeting approaches will be particularly informative (eg, for BCL2 and EZH2 inhibitors),[22,23] and will ultimately pave the way from risk-adapted to biology-directed treatment algorithms in FL. Other potentially targetable candidate genes captured by the m7-FLIPI include the acetyltransferases *EP300* and the structurally and functionally related *CREBBP*, because mutations in these genes are primarily disruptive and may sensitize tumors to histone deacetylase inhibition.[24] Likewise, N-terminally clustered mutations in *FOXO1*[25] might affect response to inhibitors of the phosphatidylinositol 3′ OH kinase (PI3K) pathway, given that FOXO transcription factors and PI3K often function as antagonists in the biology of B cells.[26] Eventually, the relative impact of individual molecular predictors will have to be adjusted to specific molecular targeting approaches; for example, *CARD11* mutations have a relatively small m7-FLIPI coefficient in the context of immunochemotherapy, but they might well increase the risk of treatment failure in patients receiving BTK inhibitors by activating NF-κB signaling downstream of BTK, as has been shown for ibrutinib for relapsed/refractory diffuse large B-cell lymphoma.[27] Several large and collaborative efforts are underway to address these questions.

In summary, the m7-FLIPI currently represents the most promising predictor for treatment outcome of patients receiving first-line immunochemotherapy, including patients with early treatment failure but not fulfilling the POD24 criteria, and should be evaluated in prospective trials of precision medicine approaches in FL.

Acknowledgments

This study was supported by the Max-Eder Program of the Deutsche Krebshilfe e.V. (110659) (O.W.), the Deutsche Forschungsgemeinschaft (DFG-SFB/CRC-1243, TP-A11) (O.W.), and in part by a Program Project Grant from the Terry Fox Research Institute (1023) (J.M.C. and R.D.G.).

G.O. receives funding from the Robert-Bosch-Foundation. D.M.W. is a Leukemia and Lymphoma Scholar.

Authorship

Contribution: O.W., D.M.W., A.D.Z., P.M.B., J.W.F., S.A., R.D.G., and W.H. contributed to study design; V.J., M.U., and O.W. performed data analysis and review, and interpretation of data; V.J. and O.W. created figures and wrote the initial manuscript; and R.K., A.M.S., M.S., H.H., M.H.D., A.R., G.O., W.K., L.H.S., J.M.C., and R.D.G. provided patient samples and data.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

The current affiliation for M.S. is Department of Internal Medicine II, Hematology Laboratory Kiel, Schleswig-Holstein University Hospital, Campus Kiel, Kiel, Germany.

Correspondence: Oliver Weigert, University Hospital of the Ludwig-Maximilians-University Munich, Medical Department III, Laboratory for Experimental Leukemia and Lymphoma Research (ELLF), Max-Lebsche Platz 30, 81377 Munich, Germany; e-mail: oliver.weigert@med.uni-muenchen.de.

References

1. Hiddemann W, Cheson BD. How we manage follicular lymphoma. *Leukemia.* 2014;28(7):1388-1395.
2. Kridel R, Sehn LH, Gascoyne RD. Pathogenesis of follicular lymphoma. *J Clin Invest.* 2012;122(10):3424-3431.
3. Mozessohn L, Cheung MC, Crump M, et al. Chemoimmunotherapy resistant follicular lymphoma: predictors of resistance, association with transformation and prognosis. *Leuk Lymphoma.* 2014;55(11):2502-2507.
4. Casulo C, Byrtek M, Dawson KL, et al. Early relapse of follicular lymphoma after rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone defines patients at high risk for death: an analysis from the National LymphoCare Study. *J Clin Oncol.* 2015;33(23):2516-2522.
5. Maurer MJ, Ghesquieres H, Ansell SM, et al. Event-free survival at 12 months (EFS12) from diagnosis is a robust endpoint for disease-related survival in patients with follicular lymphoma in the immunochemotherapy era [abstract]. *Blood.* 2014;124(21). Abstract 1664.
6. Maurer MJ, Ghesquières H, Jais JP, et al. Event-free survival at 24 months is a robust end point for disease-related outcome in diffuse large B-cell lymphoma treated with immunochemotherapy. *J Clin Oncol.* 2014;32(10):1066-1073.
7. Sargent DJ, Shi Q, De Bedout S, et al. Evaluation of complete response rate at 30 months (CR30) as a surrogate for progression-free survival (PFS) in first-line follicular lymphoma (FL) studies: results from the prospectively specified Follicular Lymphoma Analysis of Surrogacy Hypothesis (FLASH) analysis with individual patient data (IPD) of 3,837 patients (pts) [abstract]. *J Clin Oncol.* 2015;33(suppl). Abstract 8504.
8. Pastore A, Jurinovic V, Kridel R, et al. Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. *Lancet Oncol.* 2015;16(9):1111-1122.
9. Hiddemann W, Kneba M, Dreyling M, et al. Frontline therapy with rituximab added to the combination of cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) significantly improves the outcome for patients with advanced-stage follicular lymphoma compared with therapy with CHOP alone: results of a prospective randomized study of the German Low-Grade Lymphoma Study Group. *Blood.* 2005;106(12):3725-3732.
10. O'Shea D, O'Riain C, Taylor C, et al. The presence of TP53 mutation at diagnosis of follicular lymphoma identifies a high-risk group of patients with shortened time to disease progression and poorer overall survival. *Blood.* 2008;112(8):3126-3129.
11. Cheson BD, Pfistner B, Juweid ME, et al; International Harmonization Project on Lymphoma. Revised response criteria for malignant lymphoma. *J Clin Oncol.* 2007;25(5):579-586.
12. Salles G, Seymour JF, Offner F, et al. Rituximab maintenance for 2 years in patients with high tumour burden follicular lymphoma responding to rituximab plus chemotherapy (PRIMA): a phase 3, randomised controlled trial. *Lancet.* 2011;377(9759):42-51.
13. Rambaldi A, Lazzari M, Manzoni C, et al. Monitoring of minimal residual disease after CHOP and rituximab in previously untreated patients with follicular lymphoma. *Blood.* 2002;99(3):856-862.
14. Galimberti S, Luminari S, Ciabatti E, et al. Minimal residual disease after conventional treatment significantly impacts on progression-free survival of patients with follicular lymphoma: the FIL FOLL05 trial. *Clin Cancer Res.* 2014;20(24):6398-6405.
15. Pott C, Brüggemann M, Ritgen M, et al. MRD detection in B-cell non-Hodgkin lymphomas using Ig gene rearrangements and chromosomal translocations as targets for real-time quantitative PCR. *Methods Mol Biol.* 2012;971:175-200.
16. Ladetto M, Bruggemann M, Monitillo L, et al. Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders. *Leukemia.* 2013;28(6):1299-1307.
17. Roschewski M, Dunleavy K, Pittaluga S, et al. Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B-cell lymphoma: a correlative biomarker study. *Lancet Oncol.* 2015;16(5):541-549.
18. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *Br J Cancer.* 2014;110(3):551-555.
19. Federico M, Bellei M, Marcheselli L, et al. Follicular lymphoma international prognostic index 2: a new prognostic index for follicular lymphoma developed by the international follicular lymphoma prognostic factor project. *J Clin Oncol.* 2009;27(27):4555-4562.
20. Rummel MJ, Niederle N, Maschmeyer G, et al; Study group indolent Lymphomas (StiL). Bendamustine plus rituximab versus CHOP plus rituximab as first-line treatment for patients with indolent and mantle-cell lymphomas: an open-label, multicentre, randomised, phase 3 non-inferiority trial. *Lancet.* 2013;381(9873):1203-1210.
21. Flinn IW, van der Jagt R, Kahl BS, et al. Randomized trial of bendamustine-rituximab or R-CHOP/R-CVP in first-line treatment of indolent NHL or MCL: the BRIGHT study. *Blood.* 2014;123(19):2944-2952.
22. Wilson WH, O'Connor OA, Czuczman MS, et al. Navitoclax, a targeted high-affinity inhibitor of BCL-2, in lymphoid malignancies: a phase 1 dose-escalation study of safety, pharmacokinetics, pharmacodynamics, and antitumour activity. *Lancet Oncol.* 2010;11(12):1149-1159.
23. McCabe MT, Ott HM, Ganji G, et al. EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature.* 2012;492(7427):108-112.
24. Andersen CL, Asmar F, Klausen T, Hasselbalch H, Grønbæk K. Somatic mutations of the CREBBP and EP300 genes affect response to histone deacetylase inhibition in malignant DLBCL clones. *Leuk Res Rep.* 2012;2(1):1-3.
25. Trinh DL, Scott DW, Morin RD, et al. Analysis of FOXO1 mutations in diffuse large B-cell lymphoma. *Blood.* 2013;121(18):3666-3674.
26. Sander S, Chu VT, Yasuda T, et al. PI3 Kinase and FOXO1 Transcription Factor Activity Differentially Control B Cells in the Germinal Center Light and Dark Zones. *Immunity.* 2015;43(6):1075-1086.
27. Wilson WH, Young RM, Schmitz R, et al. Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. *Nat Med.* 2015;21(8):922-926.

3.3 Publication 3 - *Leukemia*, 2011

Herold T, **Jurinovic V**, Metzeler KH, et al. An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. *Leukemia*, 25(10):1639-1645, 2011.

constitutive cytokine stimulation induces a biased engraftment of a susceptible sub-population of human leukemic cells.

We conclude that with more than 81% of engraftment, non-irradiated NSG mice are an excellent tool for xenotransplantation using unsorted primary AML samples from both children and adults. Serial transplantations with hCD45+ cells of as low as 4×10^3 cells using intrafemoral injection can be successfully applied and maintain cytogenetic stability from different sources with a reduced median time to engraftment. PCR amplification of the α-satellite region of human chromosome 17 proved to be simple and highly predictive for engraftment screening of human cells. The NSG model is a powerful tool for a variety of requirements like amplification of primary human AML samples, LS-IC assays and to assess individualized molecular treatment modalities *in vivo* on a large scale.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank the members of the laboratory for Cytogenetic and Molecular Genetics at the Department of Internal Medicine III for their support with the primary adult samples, the AML–BFM group for providing the clinical information and patient data and all members of the animal facility of the University of Ulm. The technical assistance of Beate Junk is greatly appreciated. This work was supported by a grant from the Deutsche José Carreras Leukämie-Stiftung e.V. (DJCLS R07/23) to S Corbacioglu. All animal experiments were conducted according to 54-2532.1-27/09 approved protocol.

M Malaisé[1,7], M Neumeier[1,7], C Botteron[1], K Döhner[2], D Reinhardt[3], B Schlegelberger[4], G Göhring[4], B Gruhn[5], K-M Debatin[6] and S Corbacioglu[1]
[1]*Department of Pediatrics, University of Regensburg, Regensburg, Germany;*
[2]*Department of Internal Medicine III, University Hospital of Ulm, Ulm, Germany;*
[3]*Department of Pediatrics, University of Hannover, Hannover, Germany;*
[4]*Institute of Cell and Molecular Pathology, Hannover Medical School, Hannover, Germany;*
[5]*Department of Pediatrics, University Children's Hospital Jena, University of Jena, Jena, Germany and*
[6]*Department of Pediatrics, University Hospital of Ulm, Ulm, Germany*
E-mail: selim.corbacioglu@klinik.uni-regensburg.de
[7]*These authors contributed equally to this work.*

References

1 Shultz LD, Pearson T, King M, Giassi L, Carney L, Gott B *et al.* Humanized NOD/LtSz-scid IL2 receptor common gamma chain knockout mice in diabetes research. *Ann NY Acad Sci* 2007; **1103**: 77–89.
2 Shultz LD, Lyons BL, Burzenski LM, Gott B, Chen X, Chaleff S *et al.* Human lymphoid and myeloid cell development in NOD/LtSz-scid IL2R gamma null mice engrafted with mobilized human hemopoietic stem cells. *J Immunol* 2005; **174**: 6477–6489.
3 Ishikawa F, Yoshida S, Saito Y, Hijikata A, Kitamura H, Tanaka S *et al.* Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region. *Nat Biotechnol* 2007; **25**: 1315–1321.
4 Agliano A, Martin-Padura I, Mancuso P, Marighetti P, Rabascio C, Pruneri G *et al.* Human acute leukemia cells injected in NOD/LtSz-scid/IL-2Rgamma null mice generate a faster and more efficient disease compared to other NOD/scid-related strains. *Int J Cancer* 2008; **123**: 2222–2227.
5 Sanchez PV, Perry RL, Sarry JE, Perl AE, Murphy K, Swider CR *et al.* A robust xenotransplantation model for acute myeloid leukemia. *Leukemia* 2009; **23**: 2109–2117.
6 Willasch AM, Gruhn B, Coliva T, Kalinova M, Schneider G, Kreyenberg H *et al.* Standardization of WT1 mRNA quantitation for minimal residual disease monitoring in childhood AML and implications of WT1 gene mutations: a European multicenter study. *Leukemia* 2009; **23**: 1472–1479.
7 Wunderlich M, Chou FS, Link KA, Mizukawa B, Perry RL, Carroll M *et al.* AML xenograft efficiency is significantly improved in NOD/SCID-IL2RG mice constitutively expressing human SCF, GM-CSF and IL-3. *Leukemia* 2010; **24**: 1785–1788.
8 Nicolini FE, Cashman JD, Hogge DE, Humphries RK, Eaves CJ. NOD/SCID mice engineered to express human IL-3, GM-CSF and Steel factor constitutively mobilize engrafted human progenitors and compromise human stem cell regeneration. *Leukemia* 2004; **18**: 341–347.

Supplementary Information accompanies the paper on the Leukemia website (http://www.nature.com/leu)

An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia

Leukemia (2011) **25,** 1639–1645; doi:10.1038/leu.2011.125; published online 31 May 2011

The clinical course of chronic lymphocytic leukemia (CLL) is highly variable, ranging from slow progression and survival for several decades to rapidly progressive and chemotherapy-resistant disease with death within 1 year of diagnosis. The hierarchical model of common genomic aberrations determined by interphase fluorescence *in situ* hybridization (FISH) and the analysis of the mutational status of the immunoglobulin heavy-chain variable region genes (*IGVH* status) are broadly used molecular markers to predict the prognosis of CLL patients. Despite the high prognostic value of *IGVH* status and FISH analysis, the clinical course of some patients defies the predictions.[1] Until now the presence of del(17p) or a TP53 mutation are the only established predictive markers for therapy response.[2] However, they are infrequent in newly diagnosed patients (del(17p): 4%; additional TP53 mutation: 1.1%),[3,4] and sometimes misleading (overall survival (OS) rate 65% at 3 years in treatment-naïve patients).[1] Thus, a more accurate easy assessable risk classifier for CLL patients is desirable.

To improve our ability to predict the prognosis of CLL patients, gene expression profiling (GEP) and microRNA expression levels were used to develop prognostic scores.[5,6]

Calin *et al*.[6] reported a molecular signature of 13 microRNAs associated with the expression of *ZAP70,* the mutational status of *IGVH,* and the time between diagnosis and initial treatment. Rodriguez *et al*.[5] used a small custom oligonucleotide microarray to generate a prognostic model based on the expression of seven genes (including genes involved in *Wnt* and *NF-κB* signaling) for time to treatment (TTT) and validated this model in an independent CLL series. Although these studies provided new insights into the biology of CLL, all these signatures fell short of surpassing the traditional genetic markers in prognostic power; they were generated on small micro array platforms or were not validated using routine diagnostic techniques like real-time PCR (RT-PCR) or—most importantly—they were not shown to predict overall survival.

Therefore, we set out to develop and validate a simple but powerful gene expression score (PS.8) that predicts survival in CLL patients by correlating the survival data of a large patient cohort with genome-wide gene expression data.

Peripheral blood (PB) or bone marrow (BM) samples from newly diagnosed CLL patients ($N=124$), patients with preexisting CLL ($N=171$: untreated = 92; pretreated = 58; unknown = 21), or unknown disease stage ($N=5$) were analyzed in the Laboratory for Leukemia Diagnostics, Department of Internal Medicine III, University of Munich, Germany, between 2001 and 2007. The samples were received for routine diagnostics from hospitals and private practitioners in Germany. The diagnostic work-up included standard cytomorphology, multiparameter immunophenotyping, interphase FISH (11q22.3, 12cen, 13q14.3, 17p13.1 and IgH locus) and *IGVH* mutational analysis. A total of 149 patients were analyzed by oligonucleotide microarrays (44754 probe sets, Affymetrix, Santa Clara, CA, USA) and served as training set (accrued from 2001 to 2005). Two patients were analyzed twice at different disease stages (151 microarrays from 149 patients). These two patients received therapy between their two analysis time points and were analyzed the second time at relapse. There were 2.5 and 3 years time gaps between these analysis time points. We chose to treat the two analysis results from each of these two patients as separate data points because of the disease evolution that had taken place. The validation cohort consisted of 149 additional patients (accrued from 2005 to 2007). Follow-up data from the time of diagnostic work-up (study entry) were obtained from the Munich Cancer Registry of the Munich Cancer Center or our clinical database. The patients from both cohorts (training set and validation set) were not treated on a specific clinical trial. Patient inclusion in the training and validation set of our study was only based on sample availability.

As we expected that gene expression profiles change over time as the disease progresses, we defined the time of molecular assessment as the starting time point for OS and TTT. Patient characteristics are shown in Table 1 and in more detail in Supplementary Table 1. Patient data were anonymized before analysis. The retrospective study design was approved by the institutional review board of the medical faculty of the University of Munich.

The details of the microarray analysis and the development process of PS.8 are given in the supplement (see also Flow chart; Supplementary Figure 1). In brief, we used a microarray data set of 151 CLL samples (44 Affymetrix HG-U133 A&B and 107 Affymetrix HG-U133 Plus 2.0 chips) to identify genes associated

Table 1 Patient characteristics

Variable	*Training set*	*Validation set*	P-*value*
No. of included patients	151[a]	149	
Median age, years (range)	63 (30–84)	63 (33–85)	0.73
Male sex, no. (%)	89 (58.9)	99 (66.4)	0.19
Evaluable for FISH, no (%)	151 (100)	147 (98.7)	
Del. 17p13, no. (%)	13 (8.6)	7 (4.8)	0.25
Del. 11q22–23, no. (%)	18 (12)	15 (10.2)	0.71
Trisomy 12q13, no (%)	20 (13.2)	16 (10.9)	0.6
'Normal' FISH, no (%)	36 (24)	44 (29.9)	0.3
Del. 13q14, no (%)	87 (57.6)	86 (58.5)	0.91
Del. 13q14 (single), no (%)	63 (42)	67 (45.6)	0.56
Evaluable for IGVH-status, no. (%)	134 (88.7)	133 (89.3)	
IGVH mutated, no. (%)	68 (50.7)	66 (49.6)	0.9
IGVH unmutated, no. (%)	66 (49.3)	67 (50.4)	0.9
Median *IGVH* homology (%)	97.8	98.2	0.38
VH3-21 (%)	8 (6.1)	18 (14)	0.04
Patients evaluable for Binet stage, no.	106	101	
Binet A at enrollment, no. (%)	58 (54.7)	52 (51.5)	0.68
Binet B at enrollment, no. (%)	23 (21.7)	25 (24.8)	0.62
Binet C at enrollment, no. (%)	25 (23.6)	24 (23.8)	1
Median white-cell count, cells/mm^3	35 500	30 000	0.11
Median hemoglobin, g/dl	13.6	13.9	0.85
Median platelet count, cells/mm^3	165 500	161 000	0.92
No. of newly diagnosed CLL (%)	65 (44.2)	59 (39.9)	0.48
No. of preexisting untreated CLL (%)	57 (38.8)	56 (37.8)	0.48
Relapsed CLL, no. (%)	25 (17)	33 (22.3)	0.3
Samples evaluable for TTT, no.	101	105	
Treated at progression, no. (%)	56 (55.4)	57 (54.3)	0.89
Samples evaluable for OS, no.	151	149	
Deceased, no. (%)	41 (27.2)	41 (27.5)	
Median follow-up, years (range)	4.1 (0–7)	4.2 (0–5)	0.31

Abbreviations: CLL, chronic lymphocytic leukemia; Del., deletion; FISH, fluorescence *in situ* hybridization; OS, overall survival.
[a]Note that 2 patients were analyzed twice at different disease stages and were considered as different samples due to disease evolution; the data therefore consist of 149 patients with 151 microarrays.

with OS. Then, a prognostic score (PS.8) was constructed based on the weighted expression levels of eight selected genes that performed best in the microarray and quantitative RT-PCR (qRT-PCR) assays following the procedure proposed by Bair and Tibshirani (supervised principal components approach). For validation, we measured the expression levels of the selected genes with qRT-PCR in 149 independent CLL patients. Univariate and multivariate Cox regressions were used to assess the effect of the score on survival. Prediction error curves were used to assess prediction quality. The microarray data set is accessible at Gene Expression Omnibus (GSE22762).

The training and validation set had comparable distributions of genomic aberrations, the *IGVH* status, median white-cell and platelet counts, median hemoglobin levels, age, Binet stage, number of pretreatments, number and schedules of treatments and the median follow-up time. Significantly more patients with a VH3-21 rearrangement were present in the validation set and different pretreatment schedules were used in the groups (Table 1 and Supplementary Table 1). The prognostic eight gene based expression score (PS.8) was derived from the training set (Table 2) using the procedures described above (see also Flow chart; Supplementary Figure 1). PS.8 values of the 151 training microarrays were divided into three groups (high-, intermediate-, poor-risk group) using cutoffs at the 20th and 80th percentiles. Figure 1 demonstrates that PS.8 performs well—as expected—in the training data set. It is highly predictive ($P<0.001$) both for OS (Figure 1a) and TTT (Figure 1b). Additionally, using publicly available small, independent microarray data sets from CLL patients (GSE12734 from 14 patients and GSE4392 from 16 patients), we could show that our prognostic score performed surprisingly well in the classification of predefined risk groups (Supplementary Figure 2).

Quantitative RT-PCR for the eight genes of PS.8 was performed with cDNA samples from 149 independent CLL patients. The score was calculated using the weights derived from the training data. PS.8 predicted OS ($P<0.001$) and TTT ($P<0.001$) in the validation cohort. Representative Kaplan–Meier estimates for OS and TTT in patients with high (>80 percentile), low ($<20\%$ percentile) and intermediate score values are shown in Figures 1c and d. Furthermore, PS.8 predicted OS ($P=0.002$) and TTT ($P<0.001$) in patients without previous treatments (Figures 1e and f). The defined risk groups showed significant associations to markers of tumor burden (white blood cell count, Binet stage) and *IGVH* status and the deletion of 11q but not to other common genomic markers detected by FISH (Supplementary Table 2).

The performance of PS.8 was additionally evaluated in patient subgroups defined by known risk factors. Patients with mutated *IGVH* genes and no 17p13 or 11q23 deletions on FISH analysis represent patients with a favorable prognosis, whereas patients with unmutated *IGVH* genes or with a 17p13 or an 11q23 deletion have an unfavorable prognosis. PS.8 was able to significantly predict OS for the patients in the unfavorable prognosis group ($P=0.01$), but not in the favorable prognosis group ($P=0.2$). The latter result is most probably due to the weak power of the test caused by a low event rate (15%) in this group. Representative Kaplan–Meier curves obtained by dichotomization of the score are shown in Figures 2a and c. TTT was significantly predicted by PS.8 for patients with unfavorable prognosis ($P<0.001$) as well as for patients with favorable prognosis ($P=0.009$). Representative Kaplan–Meier curves are shown in Figures 2b and d. A total of 52 patients from our validation data set were defined as Binet A at time of first assessment. The mere four events in this group made comparisons of OS impossible. However, the analysis for TTT showed a significant effect of PS.8 in Binet A patients ($P=0.01$). Representative Kaplan–Meier curves are shown in Figure 2e. We did not analyze Binet stage B and C patients because these subgroups were too small. PS.8 was also predictive for the 59 newly diagnosed patients (OS $P=0.02$; TTT $P=0.003$). A representative Kaplan–Meier curve for OS in this group is shown in Figure 2f. Even though PS.8 was developed using gene expression data from unsorted PB mononuclear cells, the score was validated and performed well on expression data obtained from BM or PB irrespective of whether the cells had been sorted for CD19 positivity or not. Detailed results of univariate Cox regression for all subgroup analysis are given in the Supplementary Table 3.

In the validation set, multivariate Cox regression models were fitted to OS and TTT using PS.8 (continuous), *IGVH* status, 17p13 deletion, 11q22–23 deletion, age (<65 years vs $\geqslant 65$ years) and sex as covariates. Binet stage was not included as covariate because of incomplete data. PS.8 had a highly significant association with both endpoints (Table 3). For OS, 17p13 deletion, age and PS.8 were the only significant covariates. In the multivariate model for TTT, PS.8 is dominating all other covariates and is the only covariate significantly associated with TTT. Of note, the hazard ratio of PS.8 is given for one unit of change of the score (range of PS.8: -0.4 to $+5.5$). For example, an increase in the score from 1.5 to 2.5 results in a 1.92-fold increase in the risk for death (confidence interval (CI): 1.39–2.65). An increase of PS.8 from 1.5 to 3.5 increases the risk by 1.92^2 ($=3.69$) and so on (CI: 1.93–7.02).

Table 2 Genes contained in the eight gene prognostic score

Gene		*Genomic location*	*Affymetrix probe set*	*ABI assay*[a]	*Weight*	*Regulation in poor prognosis*
SFTPB	Surfactant protein B	2p11.2	37004_at	Hs01090658_g1	0.16	Up
MGAT4A	Mannosyl glycoprotein beta acetylglucosaminyltransferase	2q12	226039_at	Hs00923405_m1	−0.151	Down
TCF7	Transcription factor 7	5q31.1	205255_x_at	Hs00175273_m1	−0.096	Down
MGC29506	Proapoptotic caspase adapter protein precursor	5q31.2	221286_s_at	Hs00414907_m1	0.089	Up
PLEKHA1	Pleckstrin homology domain containing, family A member 1	10q26.13	226247_at	Hs00608662_m1	−0.11	Down
PDE8A	Phosphodiesterase 8A	15q25.3	212522_at	Hs00400174_m1	−0.108	Down
MSI2	Musashi homolog 2 (*Drosophila*)	17q22	243010_at	Hs00292670_m1	0.081	Up
NRIP1	Nuclear receptor interacting protein 1	21q11.2	202600_s_at 202599_s_at	Hs00942766_s1	−0.208	Down

[a]Applied Biosystems low density array (LDA) primer.

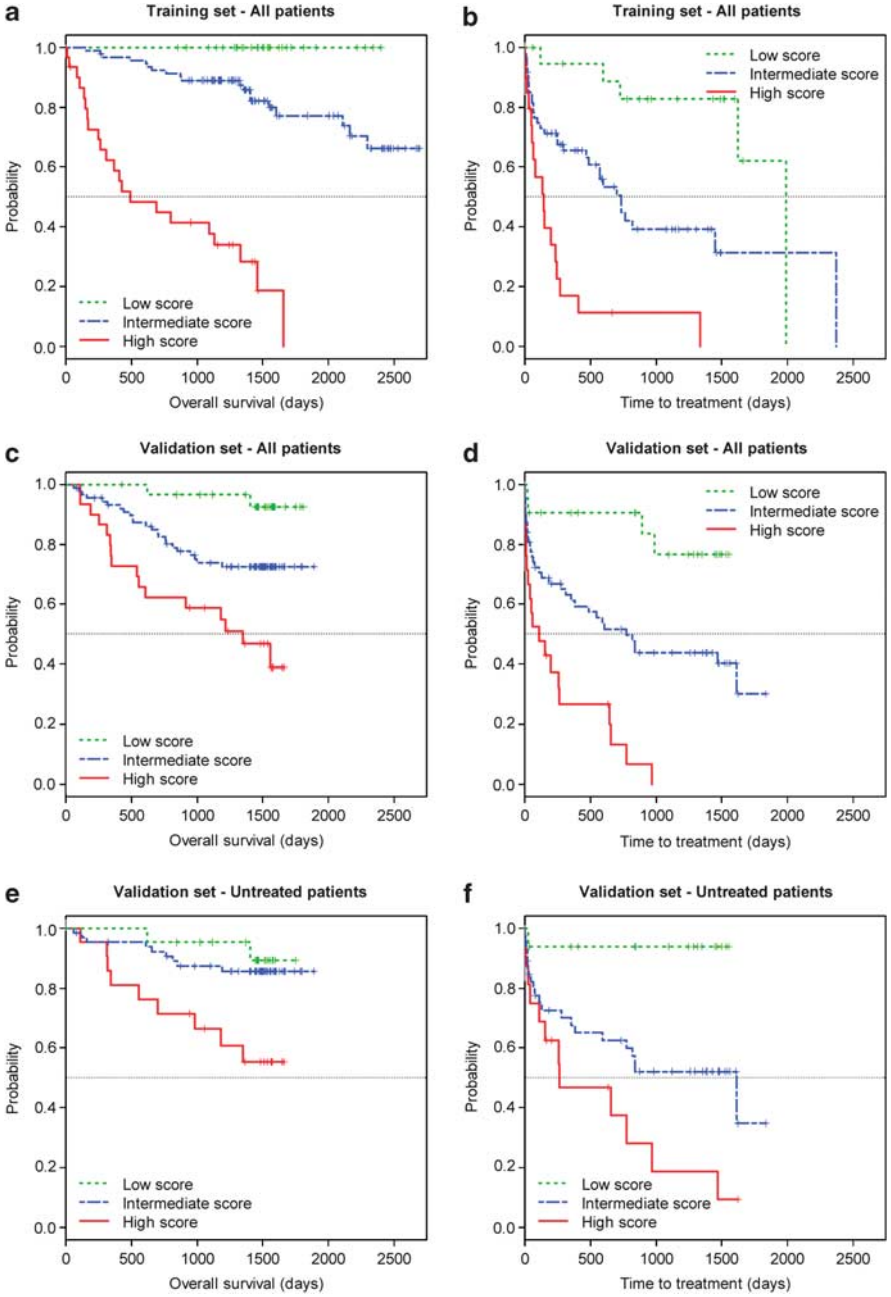

Figure 1 OS and TTT in training and validation set. Prediction of OS and TTT in the training set (**a**,**b**; microarray group) and validation set (**c**,**d**; qRT-PCR group). (**e**,**f**) Prediction of OS and TTT in patients without previous treatment in the validation set. To display the linear score a cutoff of 20% of the highest and lowest scores was chosen (log-rank test: **a**–**d**, **f**: $P<0.001$, (**e**) $P=0.004$; median OS and TTT in the low/intermediate/high risk groups: training set not reached (NR)/NR/16 months for OS and 66/24/5 months for TTT, validation set NR/NR/45 months for OS and NR/26/4 months for TTT, validation set only untreated patients NR/NR/NR for OS and NR/54/9 months for TTT).

Subsequently, we excluded all pretreated patients and patients whose treatment status at analysis was unknown from these models. In this analysis, age, 17p-deletion and PS.8 again remained as the only significant covariates for OS, and PS.8 was the only significant covariate for TTT (Supplementary Table 4a). In a multivariate model with PS.8 as categorized variable (using the cutoffs from Figure 1) fitted to OS and TTT, PS.8 showed very similar results compared with the analysis of PS.8 as a continuous variable (Supplementary Table 4b). To asses whether PS.8 is only a surrogate for tumor mass, we included Binet stage and PS.8 in an additional multivariate model (Supplementary Table 5a). In this model PS.8 remains as significant covariate for OS and TTT. This is also the case for patients without previous treatments (Supplementary Table 5b).

The additional predictive value of PS.8 was assessed using prediction error curves. For OS, a model solely based on PS.8 was superior to FISH and *IGVH* status as single markers and similar to the combined model of FISH and the *IGVH* status (Supplementary Figure 3a). The model incorporating PS.8 as well as FISH and *IGVH* status performed best. For TTT, PS.8 was superior to the single parameters and also to a combined FISH and *IGVH* status model. The addition of FISH and *IGVH* status did not increase the performance of the prognostic score (Supplementary Figure 3b).

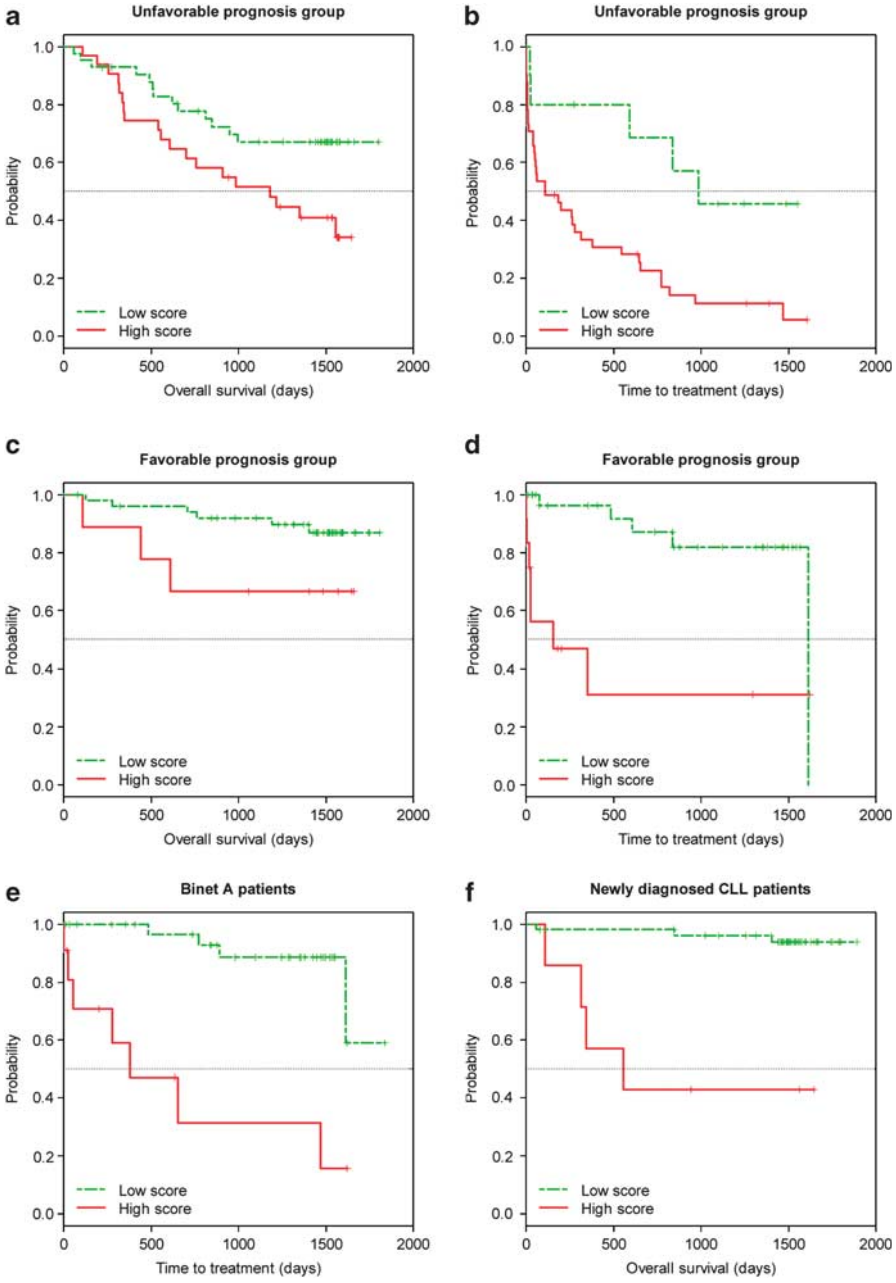

Figure 2 OS and TTT in subgroups of the validation set. Prediction of OS (**a**,**c**,**f**) and TTT (**b**,**d**,**e**) in the unfavorable prognosis subgroup (unmutated *IGVH* genes or 17p13 or 11q23 deletion; **a**,**b**), favorable prognosis subgroup (mutated *IGVH* genes, no 17p13 or 11q23 deletions; **c**,**d**), Binet A patient group (**e**) and newly diagnosed CLL patients (**f**) in the validation set. A dichotomizing cutoff corresponding in size to the number of events in the subgroups was used (**a**: 32/74, 43.2%; **b**: 41/51, 80.4%; **c**: 9/60, 15%; **d**: 12/42, 28.6%; **e**: 11/45, 24.4%; **f**: 7/59, 11.9%); Log-rank test: (**a**): $P=0.03$; (**b**): $P=0.008$; (**c**): $P=0.07$; (**d–f**): $P<0.001$.

In summary, we correlated genome-wide micro array derived gene expression values with OS in a large and heterogeneous group of CLL patients. This approach resulted in the development of a score (PS.8) based on the expression levels of eight genes. The score was validated on a different technical platform (qRT-PCR) in an independent group of patients. Importantly, PS.8 showed additional prognostic value for OS and TTT compared with the established genetic markers. Specifically, PS.8 was able to add information in several subgroups defined by the established molecular markers and in Binet A patients. The analysis of the predictive performance using prediction error curves yielded superior performances for the models containing the prognostic score compared with the models based on FISH and *IGVH* status only. PS.8 was highly significant in the multivariate analysis of previously untreated patients. Despite the heterogeneous validation group and the time difference in sampling, the gene expression score could be validated in an independent patient cohort. PS.8 remained a prognostic marker in a multivariate analysis, which included the most powerful prognostic markers in CLL (FISH and *IGVH*). These data strongly indicated that PS.8 is a highly significant and valid risk predictor.

Several of the genes contained in PS.8 are likely to have an important role in the pathogenesis of CLL. For example, low expression levels of *TCF7* (T cell specific, HMG box), a

Table 3 Multivariate analysis in the validation set

Variable	*Overall survival*[a]		*Time to treatment*[b]	
	HR (95% CI)	P-*value*	*HR (95% CI)*	P-*value*
PS.8 (continuous)	1.92 (1.39–2.65)	<0.001	1.64 (1.27–2.11)	<0.001
IGVH unmutated	1.66 (0.75–3.65)	0.21	1.86 (0.87–3.97)	0.11
Del. 11q	1.23 (0.54–2.77)	0.62	1.43 (0.65–3.14)	0.38
Del. 17p	7.05 (2.27–21.9)	<0.001	2.64 (0.98–7.1)	0.05
Age (⩾65 years)	6.04 (2.9–12.59)	<0.001	1.53 (0.86–2.73)	0.15
Sex (♂/♀)	1.52 (0.8–2.88)	0.2	0.99 (0.54–1.84)	0.98

Abbreviations: CI, confidence interval; Del., deletion; HR, hazard ratio.
[a]N = 131/149 (87.9%); 18 observations missing due to unknown *IGVH* or FISH-status.
[b]N = 92/105 (87.6%); 13 observations missing due to unknown *IGVH* or FISH-status.
Multivariate Cox regression models with covariates PS.8 (continuous), *IGVH* status (unmutated), age (⩾65 years), sex (♂/♀), 11q22–23 and 17p13 deletion for the prediction of OS and TTT. The HR of PS.8 is given for a unit change. For example, a change of PS.8 from 1.5 to 2.5 results in a multiplication of the risk of death by 1.92 (range of PS.8: −0.4 to +5.5).

downstream target of the *Wnt* signaling pathway, were found in CLL patients with poor prognosis. Recently, Kienle *et al.*[7] also reported on the association of reduced *TCF7* expression and poor outcome in CLL. The *Wnt* signaling pathway is activated in CLL, and our data strengthen the case for an important role of *Wnt* signaling in CLL. Low transcript levels of the phosphodiesterase 8A (*PDE8A*) were associated with poor prognosis in our CLL patients. A similar association was reported by Stamatopoulos *et al.*,[8] who showed that *PDE8A* was part of a gene signature that distinguishes between *ZAP70* positive and *ZAP70* negative CLL samples. *PDE8A* might be connected to the *NF-κB* signaling pathway, which is frequently activated in lymphoid malignancies. Five genes in the score (*MSI2, PLEKHA1*, *MGC29506*, *MGAT4A*, *SFTPB*) have not been described in the context of CLL before. We show that high expression levels of *MSI2* (musashi homolog 2) in CLL are associated with poor survival. *MSI2* encodes an RNA-binding translational modulator that was recently shown to be a key regulator in the Musashi–Numb pathway and identified as a prognostic marker in chronic myelogenous leukemia.[9] *MGC29506* (*MZB1)* is an endoplasmic reticulum-localized and B cell-specific protein that was very recently shown to be a key regulator of antibody secretion, integrin activation and calcium homeostasis.[10] It should be noted that genes like *MSI2, MGC29506, MGAT4A* and *PLEKHA1* were not contained in older Affymetrix Arrays like the HGU95A chip that were used in previous studies on differential gene expression in CLL. Of note, two well-known prognostic gene expression markers in CLL, *LPL* and *ZAP70,* were not included in PS.8. It might be that the significance of *ZAP70* was weakened by the use of unselected PB mononuclear cells for the microarray data set. The *LPL* expression level was a univariate significant parameter in both data sets (data not shown), but was excluded in subsequent selection steps.

The expression levels of the genes contained in PS.8 most probably reflect the status of several important cellular pathways, for example, the *Wnt, NF-κB* and Musashi–Numb signaling pathways. It is likely that PS.8 integrates and summarizes the activity of these pathways. In this context, it should be noted that treatment regimens are on the horizons that include selective pathway inhibitors targeting, for example, *NF-κB* or *Wnt* signaling.

Taken together, we present a powerful prognostic score for OS in CLL derived from a comprehensive gene expression analysis in a large cohort of patients. This score can be determined by measuring the expression levels of eight genes and can be easily implemented in a routine diagnostic setting. Prospective trials are now required to assess the relevance of PS.8 in comparison with the established genetic markers and to evaluate the usefulness of the score to guide individualized treatment choices.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We would like to thank the MCR, MCC and IBE for cooperation. Further we would like to thank the Laboratory for Leukemia Diagnostics for providing the microarray data and patient samples and, especially, Claudia and Torsten Haferlach, Frank Dicker, Alexander Kohlmann and Susanne Schnittger. Furthermore, we would like to thank Natalia Kerber and Evelyn Zellmeier for assistance in sample processing. T Herold was supported by the KL Weigand'schen Stiftung, the Curt-Bohnewands- Fonds and the Georg und Traud Gravenhorst–Stiftung and by the Deutsche Forschungsgemeinschaft, collaborative research center SFB 684 'Molecular mechanisms of normal and malignant hematopoiesis'. This work was in part supported by grants from the German Ministry of Research and Education (BMBF; 01GS0876) and the Deutsche Forschungsgemeinschaft (SFB 684) to C. Buske and SK Bohlander. V Jurinovic was supported by the German José Carreras Leukaemia Foundation (DJCLS H06/04V) and the BMBF medical systems biology project HaematoSYS (BMBF-FKZ 03415452I).

T Herold[1], V Jurinovic[2], KH Metzeler[1], A-L Boulesteix[2], M Bergmann[1,5], T Seiler[1], M Mulaw[3], S Thoene[3], A Dufour[1], Z Pasalic[1], M Schmidberger[2], M Schmidt[4], S Schneider[1], PM Kakadia[1,3], M Feuring-Buske[5,6], J Braess[1], K Spiekermann[1,3], U Mansmann[2], W Hiddemann[1,3], C Buske[6,7] and SK Bohlander[1,3,7]

[1]*Department of Internal Medicine III, University Hospital Grosshadern, Ludwig-Maximilians-University (LMU), Munich, Germany;*
[2]*Institute for Medical Informatics, Biometry and Epidemiology (IBE), Ludwig-Maximilians-University (LMU), Munich, Germany;*
[3]*Clinical Cooperative Group Acute Leukemia, Helmholtz Center Munich for Environmental Health, Munich, Germany;*
[4]*Munich Cancer Registry (MCR) of the Munich Cancer Center (MCC) at the Institute for Medical Informatics, Biometry and Epidemiology (IBE), Ludwig-Maximilians-University (LMU), Munich, Germany;*
[5]*Department of Internal Medicine III, University Hospital Ulm, Ulm, Germany and*

[6]Institute of Experimental Cancer Research, Comprehensive Cancer Center Ulm, University of Ulm, Ulm, Germany
E-mail: stefan.bohlander@med.uni-muenchen.de or christian.buske@uni-ulm.de
[7]Joint senior authors.

References

1 Tam CS, Shanafelt TD, Wierda WG, Abruzzo LV, Van Dyke DL, O'Brien S *et al.* De novo deletion 17p13.1 chronic lymphocytic leukemia shows significant clinical heterogeneity: the M. D. Anderson and Mayo clinic experience. *Blood* 2009; **114**: 957–964.
2 Zenz T, Eichhorst B, Busch R, Denzel T, Habe S, Winkler D *et al.* TP53 mutation and survival in chronic lymphocytic leukemia. *J Clin Oncol* 2010; **28**: 4473–4479.
3 Zainuddin N, Murray F, Kanduri M, Gunnarsson R, Smedby KE, Enblad G *et al.* TP53 Mutations are infrequent in newly diagnosed chronic lymphocytic leukemia. *Leuk Res* 2011; **35**: 272–274.
4 Gunnarsson R, Isaksson A, Mansouri M, Goransson H, Jansson M, Cahill N *et al.* Large but not small copy-number alterations correlate to high-risk genomic aberrations and survival in chronic lymphocytic leukemia: a high-resolution genomic screening of newly diagnosed patients. *Leukemia* 2010; **24**: 211–215.
5 Rodriguez A, Villuendas R, Yanez L, Gomez ME, Diaz R, Pollan M *et al.* Molecular heterogeneity in chronic lymphocytic leukemia is dependent on BCR signaling: clinical correlation. *Leukemia* 2007; **21**: 1984–1991.
6 Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE *et al.* A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* 2005; **353**: 1793–1801.
7 Kienle D, Benner A, Laufle C, Winkler D, Schneider C, Buhler A *et al.* Gene expression factors as predictors of genetic risk and survival in chronic lymphocytic leukemia. *Haematologica* 2010; **95**: 102–109.
8 Stamatopoulos B, Haibe-Kains B, Equeter C, Meuleman N, Soree A, De Bruyn C *et al.* Gene expression profiling reveals differences in microenvironment interaction between patients with chronic lymphocytic leukemia expressing high versus low ZAP70 mRNA. *Haematologica* 2009; **94**: 790–799.
9 Kharas MG, Lengner CJ, Al-Shahrour F, Bullinger L, Ball B, Zaidi S *et al.* Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat Med* 2010; **16**: 903–908.
10 Flach H, Rosenbaum M, Duchniewicz M, Kim S, Zhang SL, Cahalan MD *et al.* Mzb1 protein regulates calcium homeostasis, antibody secretion, and integrin activation in innate-like B cells. *Immunity* 2010; **33**: 723–735.

Supplementary Information accompanies the paper on the Leukemia website (http://www.nature.com/leu)

A novel *ABL1* fusion to the SH2 containing inositol phosphatase-1 (*SHIP1*) in acute lymphoblastic leukemia (ALL)

Leukemia (2011) **25,** 1645–1649; doi:10.1038/leu.2011.129; published online 31 May 2011

The BCR/ABL fusion gene is the most common fusion in leukemia. The BCR/ABL fusion is the hallmark of chronic myeloid leukemia and is found in about 20% of all B-lineage acute lymphoblastic leukemias (ALL). The BCR/ABL fusion leads to the constitutive activation of the ABL1 tyrosine kinase via the oligomerization of the fusion protein mediated by the coiled-coiled domain of BCR.[1,2]

Even though the ABL tyrosine kinase is one of the most frequently activated kinases in hematological malignancy, other than *BCR*,[3] only six additional *ABL1* fusion partners have been identified: *ETV6*,[4] *RCSD1*,[5] *EML1*,[6] *NUP214*,[7] *ZMIZ1*[8] and *SFPQ*.[9] Except for *NUP214*, the other *ABL1* fusions are rare. For *RCSD1*, *EML1*, *ZMIZ1* and *SFPQ*, only case reports have been described.[10] Only the NUP214/ABL1 fusion is found recurrently in approximately 6% of the T-ALL cases.[7] Here, we describe the discovery and characterization of the novel *ABL1* fusion to the SH2-containing inositol phosphatase-1 (*SHIP1; INPP5D*).

An 18-year-old woman was diagnosed with common ALL. Standard cytogenetic analysis revealed a normal karyotype of 46,XX. The routine reverse transcription-PCR screening of the patient's bone marrow mRNA for the BCR/ABL fusion transcript resulted in the amplification of a longer-than-expected PCR product (faint band Figure 1a, lane 3, white arrow). Sequence analysis of this PCR product revealed an in-frame fusion between the 5′ portion (exon 9) of the *SHIP1* gene (*INPP5D*) and the 3′ portion of *ABL1*, starting from exon 2 (Figure 1b). The presence of the SHIP1/ABL1 fusion transcript was confirmed by reverse transcription-PCR, using a primer pair specific for the SHIP1/ABL1 fusion (Figure 1c). For a detailed description of the materials and methods used in this work, please refer to the Supplementary Material.

Fluorescence *in situ* hybridization analysis using a commercially available BCR/ABL dual color dual fusion probe (Vysis/Abbott, Wiesbaden, Germany) revealed a normal signal pattern on metaphase chromosomes (data not shown). However, about 50% of the interphase nuclei showed four ABL1 signals, suggesting an *ABL1* rearrangement (red signals in Figure 1d (I)). A dual color dual fusion SHIP1/ABL1 fluorescence *in situ* hybridization probe (Figure 1d (II)) revealed three fusion signals (yellow), in addition to a green and an orange signal for the normal *SHIP1* and *ABL1* alleles, respectively (Figure 1d (III)). To distinguish whether these fusion signals were due to the a 5′-SHIP1/3′-ABL1 or the reciprocal 5′-ABL1/3′-SHIP1 fusion, single fusion (SF) SHIP1/ABL1 and SF ABL1/SHIP1 probes were employed. The SF SHIP1/ABL1 and the SF ABL1/SHIP1 probes indicated that the cells carried two SHIP1/ABL1 fusions (Figure 1d (IV)) and one ABL1/SHIP1 fusion (Figure 1d (V)).

The new fusion partner of *ABL1*, *SHIP1* (*INPP5D*) encodes a protein of 1188 amino acids (aa; 145 kDa), with an N-terminal SH2 (scr-homology) domain, an inositol phosphatase domain and a C-terminal proline-rich region (Figure 1e). The SHIP1/ABL1 fusion protein contains the first 343 aa of SHIP1, including the SH2 domain and almost the complete ABL1 protein as it is found in other ABL1 fusion proteins (for example, in BCR/ABL1) (Figure 1e).

To determine the frequency of the SHIP1/ABL1 rearrangement in leukemia, we screened the cDNA of 63 ALL cell lines (36 pre B-ALL, 8 B-ALL and 19 T-ALL; Supplementary Table 1), using SHIP1/ABL1 fusion transcript-specific primers. None of the cell lines was positive. We expanded our search to include a series of 678 BCR/ABL-negative childhood ALL cases (age 1.4–17.4 years; mainly B-ALL), which had been examined at the cytogenetics laboratory of the University of Giessen hospital.

A. Appendix

Prognostic Signatures

July 24, 2013

1 Motivation

In the past decade, many microarray-based gene expression signatures for a wide variety of different diseases have been published. Very often, these signatures are not very stable, so that an exclusion of a few samples from the data set can sometimes result in a very different signature. This can be seen in the fact that many published gene signatures analyzing the same disease hardly overlap. Frequently, signatures contain so many genes that it is sometimes unfeasible to adopt them into everyday clinical practice. We propose to include two steps in the signature buliding process: first, the bootstrapping procedure should be used in order to stabilize the signature, followed by penalized regression that is used to shrunk the signature to a smaller size. We will show that these two steps improve the stability and produce signatures of a size that is suitable for clinical use. Moreover, our examples show that this prodecure not only improves the signature's feasibility, but it also tends to improve its performance.

2 Example data set and prognostic signature

The microarray data set we use in our example was taken from Metzeler et al. [2008]. It consists of samples from 163 adult patients with cytogenetically normal AML (CN-AML), measured on Affymetrix HG-U133A and HG-U133B chips (44754 different probe sets in total). This data set ("training set") was used by the authors to infer a probe set signature that predicts a patient's overall survival (OS). The authors used the supervised principal components method ("superpc") to construct a prognostic score (Bair and Tibshirani [2004], Bair et al. [2006]):

First, a univariate Cox regression was performed with all 44754 probe sets, and the probe sets with absolute Cox scores greater than 2.9 were used. This threshold was calculated with a 10-fold cross-validation procedure. The result was a list of 86 probe sets (66 genes) which we will call "prognostic signature".

Next, a principal component analysis was performed with only these 86 probe sets, and a correlation of each probe set with the first principal component was used as a weight for the construction of a continuous predictor. Finally, the predictor was calculated as a linear combination of expression values of the 86 probe sets multiplied with their previously obtained weights. This gives a single value for each patient, which we call a "risk score".

The risk score was tested on two independent data sets. The first validation cohort consisted of 79 CN-AML patients whose pretreatment diagnostic

evaluation was done at the Laboratory for Leukemia Diagnostics, University of Munich. The second validation cohort consisted of 64 CN-AML patients treated in the Cancer and Leukemia Group B (CALGB) 9621 study. In both cases, the risk score was shown to be significant, both in univariate and multivariate regression, the latter including already known prognostic factors age, NPM1 mutation status and FLT3 ITD/wt ratio.

Taking a closer look at this signature, we discovered that some genes occured more than once in the signature, because they were represented by more than one probe set. For example, the gene FHL1 is represented in the signature by five different highly correlated probe sets. If a gene is associated with the response and there are several probe sets representing it, it is very probable that more than one of them will end up in the signature. Thus, the relative freqency of a gene in a signature is somewhat determined by Affymetrix probe set selection. Furthermore, even if the signature finds its way into clinical practice, it is very improbable it will be measured by microarrays. For these reasons, a gene signature should always be preferred to a probe set signature: it is easier to interpret, and it can be measured in more than one way.

The training set and the first validation cohort can be downloaded from Gene Expression Omnibus (accession number GSE12417).

3 Bootstrapping improves the stability of signatures

It is well known that gene signatures from different studies analyzing the same disease have little or no overlap, meaning that the signatures are rather unstable (Ein-Dor et al. [2005]). This is probably due to the fact that many genes correlate with each other, so one gene can be substituted with another one (or a group of other genes) without changing the prognostic value of the signature. Hovewer, we will show that, even though we probably can never expect to build stable signatures out of individual genes, the stability can still be substantially improved.

We propose the following procedure: Insted of calculating a signature once on a complete data set, create n bootstrap samples and calculate a signature for each one of them with your method of choice (Buchholz et al. [2008]). For each variable, calculate the frequency with which it appears in these n signatures. Use the most frequent variables to build your final signature, with "most frequent" depending on the signature size you prefer. Here, we use k-means clustering on the vector of frequencies and define the variables in the cluster of highest frequencies to be our new signature. Since we will later use Lasso regression to shrink the signature, we want the signature size to be smaller than the sample size because Lasso has some drawbacks in the $p > n$ case (Tibshirani [1996]). On the other hand, we want the bootstrap signature to be as large as possible, so Lasso can later identify the most important genes. In our case, 3-means clustering produces the largest signatures that are still smaller than the sample size, so this is the method we decided to use.

To compare the stability of signatures with and without bootstrapping, we created 100 data sets by taking out 5% of samples ($\equiv$ 8 samples) from the original data set. For each of these data sets, we created one prognostic signature

with the superpc method ("superpc signature"), and one signature by calculating a superpc signature for 1000 bootstrap samples and aggregating them as described above ("bootstrap signatures"). The length of the superpc signatures varied from 5 to 220 (4 - 161 genes), while the length of bootstrap signatures ranged from 37 to 111 (27 - 84 genes), showing a much smaller variance. Combined, the 100 superpc signatures contained a total of 270 probe sets (195 genes), while the bootstrap signatures contained 210 (156). To compare the stability of signatures, we calculated the frequency of signature members. For example, the probe set "201397_at" from the first bootstrap signature was present in 44 other bootstrap signatures, so its frequency is 45. 29 probe sets (=13.8%) from the bootstrap signatures were present in all 100 signatures, while only 5 probe sets from the superpc signatures were present in all of them. The smoothed distribution of frequencies is shown in figure 1.

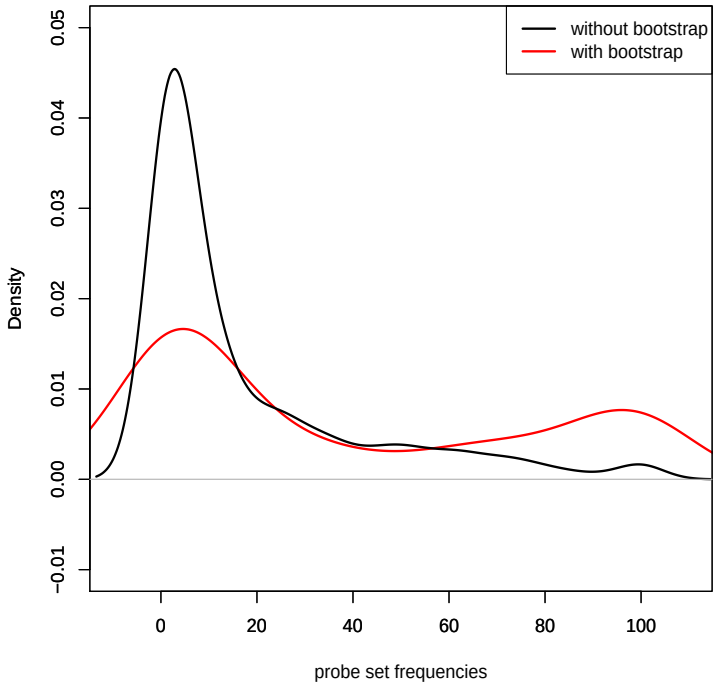

Figure 1: Smoothed frequencies of probe sets in superpc and bootstrap signatures.

To compare the performance of the signatures, we used the first validation cohort from Metzeler et al. [2008] and performed a univariate and multivariate Cox regression for each of the 200 scores. The multivariate regression included the covariables continuous risk score, FLT3 ITD, NPM1 mutation and age. The data matrix from the second validation cohort could not be used to test these scores because it was restricted in the number of probesets.

The p-values for the Cox coefficients of the risk scores are shown in figure 2.

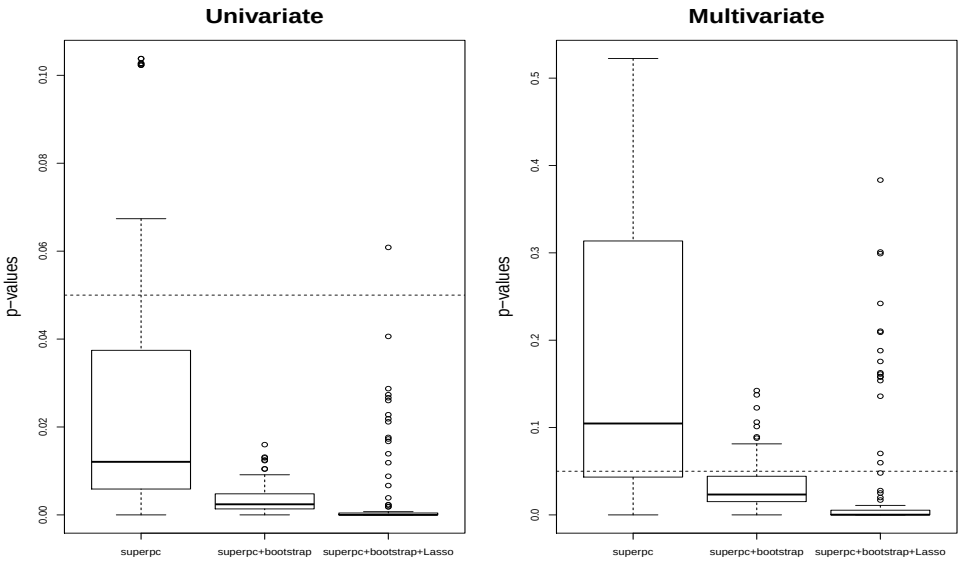

Figure 2: Summary of 100 p-values for superpc, bootstrap and Lasso signatures. Left: p-values from univariate Cox regression; Right: p-values from multivariate Cox regression with the risk score, FLT3 ITD, NPM1 and age as covariables. The dashed line represents the 0.05 significance level.

Univariately, 15 superpc scores were non-significant at the α-level of 0.05, while all bootstrap scores stayed under the significance level. Overall, bootstrapping improved the p-values in 85 cases. The picture is similar for multivariate regression, although the p-values are, on the whole, much higher. This is not surprising, since variables like mutations are mirrored in the gene expression of individuals. Examining only the association of gene expression with the response variable, we will occasionally find a gene that is significant only through its assiciation with a prognostic mutation. Putting these two variables together in a regression model, the p-value for the gene will go up compared to univariate regression. If one wants to find only those genes that are uncorrelated with known prognostic factors, one can include these factors in the signature modelling from the start.

For 2-, 4-, and 5-means clustering, the distribution of frequencies is similar to our results for 3-means clustering, with all p-values under the significance level as well. However, the median p-value increases with the increasing number of clusters (data not shown). Therefore, we recomend to first choose a signature size slightly smaller than the sample size and then to shrink it in the next step with Lasso regression. If, say, one wants a signature of 20 variables, shrinking a larger bootstrap signature with Lasso regression to the size of 20 produces much better results than simply taking the 20 most frequent variables from the pooled list of bootstrap results.

4 Penalized regression can make signatures shorter without a substantial loss of information

One problem with a large signature is its limited feasibility. The more genes we need to measure, the more time-consuming and error-prone the procedure becomes. Furthermore, a signature containing many genes is very difficult to interpret, since it is very hard to tell which of the many genes are the most interesting ones. Mostly, researchers prefer a smaller model that may have a slightly higher bias, but is easier to conduct and interpret. For this reason, penalized regression techniques are widely used in microarray analyses where researchers are dealing with many predictors. One of the most popular is Lasso regression [Tibshirani, 1996] that uses the L1-norm to shrink some regression coefficients to zero. We used Lasso to select the most important genes out of our large-sized bootstrap signatures.

We shrank 100 bootstrap signatures with the function `penalized` from the `R` library *penalized*. The optimal penalty parameter for each regression was calculated with leave-one-out cross validation. To prevent Lasso from penalizing too many variables, we set the maximal penalty parameter to 20. The new signatures were from 6 to 17 probe sets long, with a median of 10. These signatures were also tested on the first validation cohort and the p-values were compared with the p-values of complete bootstrap signatures (figure 2). 85 p-values improved by penalizing the signatures, while 14 p-values that did not improve still remained significant. Only one p-value jumped from 0.01 to 0.06 and thus missed the significance level of 0.05.

Finally, to summarize the goodness of different signatures, we calculated Nagelkerke's pseudo R^2 that reflects the improvement of the full model over the null model. We chose this particular pseudo R^2 because it was easy to compute from the output of the `R` function `coxph` and it is easier to interpret than the Cox & Snell pseudo R^2 given in the `R` output. The value was calculated for all 300 prognostic scores, and the summary for the three different methods is shown in figure 3.

5 A new signature

Using bootstrapping and Lasso regression, we developed a new signature on the training data set from Metzeler et al. [2008]. Since the bootstrap calculations are very time consuming, we did them in parallel on 100 nodes using the `R` package *snow* on our computer cluster:

```
library(snow)
cl <- makeCluster(100, type = 'MPI')
```

Like the authors, we used 10-fold cross validation to determine the optimal cutoff coefficient for each bootstrap sample. We implemented the whole procedure in the `R` function `Boot`:

```
boot.signatures <- clusterCall(cl, Boot, data.matrix, survival.days,
+                               survival.exit, 10)
```

The result is a list of 100 objects for each cluster node, each object being a list of 10 boostrap signatures. Thus, the object `boot.signatures` contains

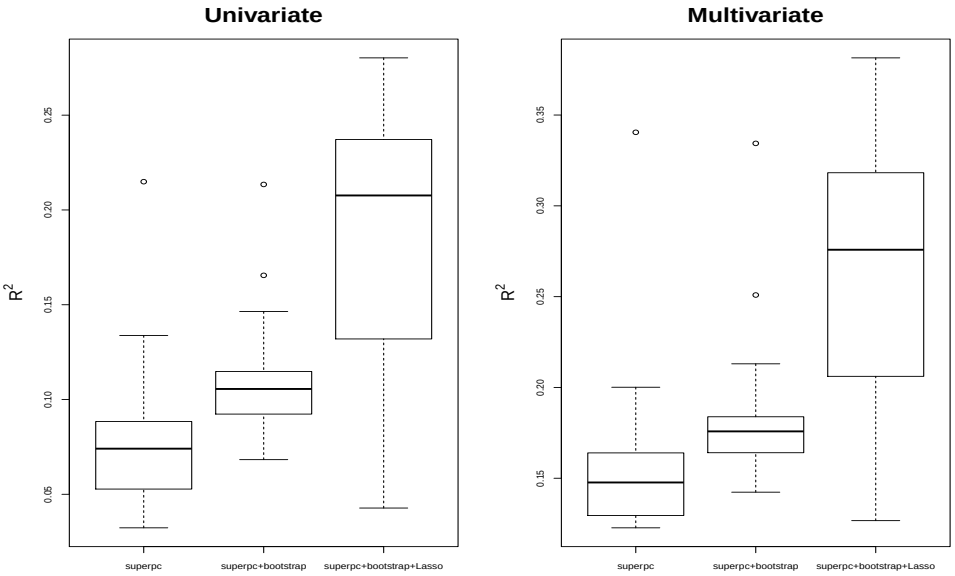

Figure 3: Summary of 100 R^2 values for superpc, bootstrap and Lasso signatures.

1000 bootstrap signatures in total. Next, we pooled all the signatures together to calculate the frequencies for every probe set:

```
probes <- unlist(boot.signatures)
freq <- table(probes)
```

Then we did 3-means clustering on the frequency vector and chose the highest cluster to be our bootstrap signature:

```
km <- kmeans(freq, 3)
signature.cluster <- which.max(km$centers)
boot.signature <- names(km$cluster)[which(km$cluster==signature.cluster)]
```

Finally, we used the package *penalized* to calculate the optimal penalty parameter and perform Lasso regression:

```
library(penalized)
opt.lambda <- optL1(Surv(survival.days, survival.exit),
+                   penalized=t(data.matrix[boot.signature, ]),
+                   maxlambda1=20)$lambda
pen <- penalized(Surv(survival.days, survival.exit),
+                penalized=t(data.matrix[boot.signature, ]),
+                lambda1=opt.lambda)
Signature <- names(pen@penalized)[which(pen@penalized!=0)]
```

The result is a prognostic signature that contains only 15 probe sets (15 genes). All the probe sets from this signature are also contained in the 86-probe

probeset	gene	15-probe set signature	86-probe set signature
203373_at	SOCS2	0.118	0.024
204419_x_at	HBG1/HBG2	-0.109	-0.013
204438_at	MRC1	0.067	0.011
208798_x_at	GOLGA8A	0.063	0.013
209386_at	TM4SF1	0.054	0.011
209760_at	KIAA0922	-0.020	-0.003
211597_s_at	HOPX	0.090	0.014
212509_s_at	MXRA7	0.031	0.005
213428_s_at	COL6A1	0.029	0.006
218086_at	NPDC1	0.059	0.012
218136_s_at	SLC25A37	-0.062	-0.01
220377_at	FAM30A	0.089	0.017
221210_s_at	NPL	-0.051	-0.009
224710_at	RAB34	0.066	0.01
241133_at	TRBV27	0.092	0.017

Table 1: List of probe sets contained in the 15-probe set signature

set signature. The probe sets and corresponding genes are shown in table 1 along with their weights in both signatures.

The risk score made of this signature was tested on the test and validation cohort and compared to the published 86-probe set risk score. Since we did not have the information on the FLT3 ITD/wt ratio for the patients from the test cohort, we included the variable FLT3 ITD in the multivariate analysis. For the exact comparison, we recalculated the multivariate analysis for the 86-probe set score including FLT3 ITD instead of FLT3 ITD/wt ratio. The coefficients and p-values for both scores and other covariables are shown in table 2 and 3.

	15-probe set signature		86-probe set signature	
Variable	HR (95% CI)	P	HR (95% CI)	P
Age	1.03 (1.00 - 1.06)	0.04	1.03 (1.00 - 1.06)	0.05
NPM1 Mutation	0.72 (0.33 - 1.53)	0.39	0.55 (0.27 - 1.15)	0.11
FLT3ITD	0.89 (0.34 - 2.34)	0.82	1.34 (0.57 - 3.15)	0.50
Risk score	2.27 (1.31 - 3.94)	0.003	1.75 (1.05 - 2.90)	0.03

Table 2: Results of multivariate Cox regression in the test cohort

	15-probe set signature		86-probe set signature	
Variable	HR (95% CI)	P	HR (95% CI)	P
Age	0.98 (0.95 - 1.01)	0.15	0.87 (0.66 - 1.13)	0.3
NPM1 Mutation	0.43 (0.18 - 1)	0.05	0.43 (0.18 - 1.01)	0.052
FLT3ITD	2.29 (0.9 - 5.84)	0.08	1.75 (0.63 - 4.81)	0.28
Risk score	2.89 (1.3 - 6.43)	0.009	3.40 (1.4 - 8.29)	0.007

Table 3: Results of multivariate Cox regression in the validation cohort

6 Conclusion

We have shown on a real data example that it is possible to shrink a large gene signature to less than 20 variables without the loss of information. Moreover, disregarding the less important genes from the signature seems to improve the specificity of the risk score.

References

Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2:511–522, 2004.

Eric Bair et al. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137, 2006.

Anika Buchholz, Norbert Holländer, and Willi Sauerbrei. On properties of predictors derived with a two-step bootstrap model averaging approach-a simulation study in the linear regression model. *Computational Statistics & Data Analysis*, 52(5):2778–2793, 2008.

Liat Ein-Dor et al. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005.

Metzeler et al. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, 112:4193–4201, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

Bibliography

[1] Anderson JR, Armitage JO, and Weisenburger DD. Epidemiology of the non-Hodgkin's lymphomas: Distributions of the major subtypes differ by geographic locations. Non-Hodgkin's Lymphoma Classification Project. *Annals of Oncology*, 9(7):717–720, 1998.

[2] Relander T, Johnson NA, Farinha P, et al. Prognostic Factors in Follicular Lymphoma. *Journal of Clinical Oncology*, 28(17):2902–2913, 2010.

[3] Kridel R, Sehn LH, and Gascoyne RD. Pathogenesis of follicular lymphoma. *The Journal of Clinical Investigation*, 122(10):3424–3431, 2012.

[4] Lister TA, Crowther D, Sutcliffe SB, et al. Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. *Journal of Clinical Oncology*, 7(11):1630–1636, 1989.

[5] Hiddemann W and Cheson BD. How we manage follicular lymphoma. *Leukemia*, 28:1388–1395, 2014.

[6] Marcus R, Imrie K, Belch A, et al. CVP chemotherapy plus rituximab compared with CVP as first-line treatment for advanced follicular lymphoma. *Blood*, 105(4):1417–1423, 2005.

[7] Herold M, Haas A, Srock S, et al. Rituximab Added to First-Line Mitoxantrone, Chlorambucil, and Prednisolone Chemotherapy Followed by Interferon Maintenance Prolongs Survival in Patients With Advanced Follicular Lymphoma: An East German Study Group Hematology and Oncology Study. *Journal of Clinical Oncology*, 25(15):1986–1992, 2007.

[8] Hiddemann W, Kneba M, Dreyling M, et al. Frontline therapy with rituximab added to the combination of cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) significantly improves the outcome for patients with advanced-stage follicular lymphoma compared with therapy with CHOP alone: results of a prospective randomized study of the German Low-grade Lymphoma Study Group. *Blood*, 106(12):3725–3732, 2005.

[9] Casulo C, Byrtek M, Dawson KL, et al. Early Relapse of Follicular Lymphoma After Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone Defines Patients at High Risk for Death: An Analysis From the National LymphoCare Study. *Journal of Clinical Oncology*, 33(23):2516–2522, 2015.

[10] Solal-Céligny P, Roy P, Colombat P, et al. Follicular Lymphoma International Prognostic Index. *Blood*, 104(5):1258–1265, 2004.

[11] van de Schans SAM, Steyerberg EW, Nijziel MR, et al. Validation, revision and extension of the Follicular Lymphoma International Prognostic Index (FLIPI) in a population-based setting. *Annals of Oncology*, 20(10):1697–1702, 2009.

[12] Buske C, Hoster E, Dreyling M, et al. The Follicular Lymphoma International Prognostic Index (FLIPI) separates high-risk from intermediate- or low-risk patients with advanced-stage follicular lymphoma treated front-line with rituximab and the combination of cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) with respect to treatment outcome. *Blood*, 108(5):1504–1508, 2006.

[13] Montoto S, López-Guillermo A, Altás A, et al. Predictive value of Follicular Lymphoma International Prognostic Index (FLIPI) in patients with follicular lymphoma at first progression. *Annals of Oncology*, 15(10):1484–1489, 2004.

[14] O'Shea D, O'Riain C, Taylor C, et al. The presence of TP53 mutation at diagnosis of follicular lymphoma identifies a high-risk group of patients with shortened time to disease progression and poorer overall survival. *Blood*, 112(8):3126–3129, 2008.

[15] Cheung KJ, Johnson NA, Affleck JG, et al. Acquired TNFRSF14 Mutations in Follicular Lymphoma Are Associated with Worse Prognosis. *Cancer Research*, 70(22):9166–9174, 2010.

[16] Correia C, Schneider PA, Dai H, et al. BCL2 mutations are associated with increased risk of transformation and shortened survival in follicular lymphoma. *Blood*, 125(4):658–667, 2015.

[17] Rozman Ciril and Montserrat E. Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 333(16):1052–1057, 1995.

[18] Binet JL, Auquier A, Dighiero G, et al. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer*, 48(1):198–206, 1981.

[19] Rai KR, Sawitsky A, Cronkite EP, et al. Clinical staging of chronic lymphocytic leukemia. *Blood*, 46(2):219–234, 1975.

[20] Hallek M, Cheson BD, Catovsky D, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute–Working Group 1996 guidelines. *Blood*, 111(12):5446–5456, 2008.

[21] Eichhorst B, Hallek M, and Dreyling M. Chronic lymphocytic leukemia: ESMO Minimum Clinical Recommendations for diagnosis, treatment and follow-up. *Annals of Oncology*, 20(suppl_4):iv102–iv104, 2009.

[22] Dighiero G and Binet JL. When and How to Treat Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 343(24):1799–1801, 2000.

[23] Damle RN, Wasil T, Fais F, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood*, 94(6):1840–1847, 1999.

[24] Hamblin TJ, Davis Z, Gardiner A, et al. Unmutated Ig V_H Genes Are Associated With a More Aggressive Form of Chronic Lymphocytic Leukemia. *Blood*, 94(6):1848–1854, 1999.

[25] Juliusson G, Oscier DG, Fitchett M, et al. Prognostic Subgroups in B-Cell Chronic Lymphocytic Leukemia Defined by Specific Chromosomal Abnormalities. *New England Journal of Medicine*, 323(11):720–724, 1990.

[26] Döhner H, Stilgenbauer S, James MR, et al. 11q Deletions Identify a New Subset of B-Cell Chronic Lymphocytic Leukemia Characterized by Extensive Nodal Involvement and Inferior Prognosis. *Blood*, 89(7):2516–2522, 1997.

[27] Dohner H, Fischer K, Bentz M, et al. p53 gene deletion predicts for poor survival and non-response to therapy with purine analogs in chronic B-cell leukemias. *Blood*, 85(6):1580–1589, 1995.

[28] Stilgenbauer S, Bullinger L, Benner A, et al. Incidence and clinical significance of 6q deletions in B cell chronic lymphocytic leukemia. *Leukemia*, 13:1331–1334, 1999.

[29] Döhner H, Stilgenbauer S, Benner A, et al. Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 343(26):1910–1916, 2000.

[30] Deuflhard P. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms.* Springer Series in Computational Mathematics (35). Springer, 2004.

[31] Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[32] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[33] Zhu J and Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443, 2004.

[34] Tibshirani R. The LASSO method for variable selection in the Cox Model. *Statistics in medicine*, 16(4):385–395, 1997.

[35] Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[36] Efron B and Tibshirani RJ. *An Introduction to the Bootstrap.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[37] Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11):559–572, 1901.

[38] Jolliffe IT. A note on the Use of Principal Components in Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):300–303, 1982.

[39] Bair E, Hastie T, Paul D, and Tibshirani R. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

Acknowledgements

I had a great fortune to have the best supervisor I can think of, professor Ulrich Mansmann, who gave me so much freedom and time to find my favorite area of research. He created an atmosphere at the IBE where I instantly felt at home, and that makes it fun to come to work every day. I am deeply thankful for his continuous support, encouragement and patience, especially in times when I had no idea where the journey was going.

My first project with Dr. Tobias Herold was so exciting for me that from that moment on, my love for hematology was born and since then has never let me go. Thanks Tobias for introducing me to the most exciting area of medicine, for being my family's personal doctor and a friend ☺

Working with Dr. Oliver Weigert has been so much fun I can hardly call it work. His excitement and enthusiasm for science are truly contagious and have been a neverending source of motivation from the first day we started working together.

I would also like to express my gratitude to Dr. Eva Hoster and Dr. Michael Unterhalt for patiently answering my many questions about everything, for the constructive criticism, and for making sure my love for chaos does not get hold of me.

Thanks to my dear friend Dr. Zlatana Pašalić for so many years of our notorious Balkanstüberl that annoyed so many people near and far, and that I miss every day since she left.

I want to thank my whole family for always being proud of me, especially my parents who always made sure I had the chance to do everything I was ever interested in. Thanks to my sister Ana for being by my side since I can remember myself.

Thanks Tamara for being in my life.