Medizinische Fakultät
der
Universität Duisburg-Essen

Aus dem Institut für Medizinische Informatik, Biometrie und
Epidemiologie

Developing a Prognostic Risk Score for Patients with Newly Diagnosed Aggressive Non-Hodgkin Lymphoma: Clinical and Statistical Issues

Inaugural–Dissertation
zur
Erlangung des Doktorgrades der
Naturwissenschaften in der Medizin
durch die Medizinische Fakultät
der Universität Duisburg-Essen

vorgelegt von
Dipl.-Stat. Jan Rekowski
aus
Düsseldorf
2019

DOI: 10.17185/duepublico/71424
URN: urn:nbn:de:hbz:464-20200608-121358-3

Dekan: Herr Univ.-Prof. Dr. med. J. Buer
1. Gutachter: Herr Univ.-Prof. Dr. rer. nat. K.-H. Jöckel
2. Gutachter: Herr Univ.-Prof. Dr. med. V. Grünwald

Tag der mündlichen Prüfung: 13. Januar 2020

Contents

1 Introduction

Non-Hodgkin lymphomas (NHL) are a heterogeneous group of diseases which are defined as any neoplasia of lymphatic cells that do not meet the criteria of Hodgkin lymphoma (Shankland et al., 2012; Armitage et al., 2017). A general distinction is between tumours stemming from B-cells and T-cells, respectively, and they can further be divided into indolent and aggressive lymphomas based on their proliferation rate. This thesis focuses on aggressive NHL only but covers both B-cell and T-cell lymphomas. While the most common entities of B-cell lymphomas are diffuse large B-cell lymphoma (DLBCL), primary mediastinal large B-cell lymphoma (PMBCL), and follicular lymphoma of grade 3 (FL), the less frequent T-cell lymphomas also contain several subtypes according to the most recent revision of the World Health Organization classification of lymphoid neoplasms (Swerdlow et al., 2016). Currently, standard therapy for all aggressive NHL is a chemotoxic regimen consisting of cyclophosphamide, doxorubicin hydrochloride, vinicristine[1], and the steroid prednisolone (CHOP; Fisher et al., 1993). Since its extension with the monoclonal antibody rituximab, the drug combination is abbreviated as R-CHOP (Michallet and Coiffier, 2009). Note, however, that rituximab is only applied in patients positive for the cluster of differentiation molecule 20 (CD20) which applies to literally all lymphomas with B-cell and only to very few with T-cell origin. Thus, the treatment protocol with the optional application of rituximab is referred to as (R-)CHOP in this thesis. Therapy results of (R-)CHOP in aggressive NHL differ by the origin of the tumour — with higher cure rates in B-cell than in T-cell lymphomas (Pfreundschuh et al., 2008; Ellin et al., 2014). However, treatment success is likewise dependent on other features such that, with the advent of personalised medicine, therapy optimisation on individual patient level also became appealing in aggressive NHL. That means, guidance is needed how to identify and treat patients who do not respond to (R-)CHOP therapy satisfactorily. The International Prognostic Index (IPI; Shipp et al., 1993) was one of the first coordinated attempts to define prognostic risk groups in terms of overall survival for NHL patients based on pre-treatment patient characteristics. It considers five risk factors and has been shown to be predictive for overall survival in aggressive NHL, and this feature has also been maintained in the rituximab era (Ziepert et al., 2010). Still, it has never been considered in guidelines to systematically inform treatment decisions. In his methodological work on updating prognostic survival models, van Houwelingen (2000)

[1]Vincristine is among others marketed under a brandname that starts with the letter "O"; therefore the abbreviation CHOP for the entire therapy regimen.

re-investigates the IPI and reviews its validity using data from the Netherlands (Hermans et al., 1995). The major point of his criticism is the weighting of the risk factors in the IPI that penalises each unfavourable characteristic with one point. In the meantime, several extensions to and modifications of the IPI have been proposed. These attempts include among others an age-adjusted IPI in the original article (Shipp et al., 1993) as well as an IPI adjusted for the lactate dehydrogenase level (Park et al., 2014), a re-distribution of the initial risk groups tailored to the rituximab era (Sehn et al., 2007), and several versions for subgroups of aggressive NHL, e.g., a version for DLBCL that improves the use of the extranodal manifestation factor (Zhou et al., 2014). However, none of these approaches includes early response to therapy. While response to therapy in oncology is traditionally assessed by "anatomic imaging alone using standard WHO [World Health Organization], RECIST [Response evaluation criteria in solid tumors], and RECIST 1.1 criteria" (Wahl et al., 2009), it can also be based on the results from imaging procedures that measure the metabolic activity of the tumour. In aggressive NHL, combined use of positron emission tomography and computed tomography (PET/CT) — usually using the radioactive tracer fluorodeoxyglucose (^{18}F-FDG) — allows to inform about the metabolic activity of a lesion, and this technique has recently become part of "standard staging for FDG-avid lymphomas" (Cheson et al., 2014). The two major competing methods are visual assessment using the qualitative Deauville criteria (Meignan et al., 2009) and the semi-quantitative deltaSUVmax approach (Lin et al., 2007) that focuses on measuring the maximum standardised uptake value (SUV) in the most active lymphoma manifestation. The Positron Emission Tomography–Guided Therapy of Aggressive Non-Hodgkin Lymphomas (PETAL) trial (Dührsen et al., 2018) was the first large-scale pivotal study to investigate treatment options for bad prognosis aggressive NHL patients where patients with unfavourable prognosis were selected according to their relative reduction of SUV between before and after the first two cycles of (R-)CHOP chemotherapy (ΔSUV).

The combination of PET/CT therapy guidance as utilised in the PETAL trial and clinical pre-treatment characteristics as, for example, considered in the IPI is from a statistical perspective equivalent to the development of a prognostic model with regard to overall survival under (R-)CHOP treatment. Prognostic models overcome the major drawback of the IPI identified by van Houwelingen (2000) and allow for estimation of individual survival probabilities given a pre-specified candidate set of variables with presumably prognostic value (Altman, 2009). Two series of scientific work have been published that deal with prognostic research in general and focus on frameworks to

develop and validate prognostic models (Steyerberg et al., 2013; Moons et al., 2009). For time-to-event endpoints like overall survival, such models can be built from many statistical methods. Beyond the Cox proportional hazards regression model (Cox, 1972) that has become standard for the analysis of right-censored data, many other techniques may be used. A non-exhaustive list of such approaches and corresponding references would, for example, include parametric accelerated failure time models (Collett, 2014, Chapter 6), spline functions (Harrell et al., 1988; Gray, 1992), fractional polynomial models (Royston and Sauerbrei, 2008), and Bayesian methods (Ibrahim et al., 2001) as well as machine learning techniques (Obermeyer and Emanuel, 2016). Performance of resulting models can be compared in various ways depending on their intended application. For an overview of performance measures for prognostic models, refer, for example, to Gerds et al. (2008) who among other concepts introduce discrimination and calibration. Briefly, discrimination considers the ability of a model to distinguish between patients with bad and good prognosis, respectively, while calibration focuses on the "agreement between observed outcomes and predictions" (Austin and Steyerberg, 2012).

The intention of this thesis is to combine early response to therapy with clinical pre-treatment characteristics. It compares the IPI and an equally simple associated risk score with three risk scores that are elicited from prognostic modelling, that is, using logistic regression (see, e.g., Hosmer et al., 2013), Cox proportional hazards regression (Cox, 1972), and the multivariable fractional polynomial time (MFPT) approach (Sauerbrei et al., 2007). The primary hypothesis of this thesis focuses on the clinical question whether a prognostic model can be developed from the PETAL data that includes early response to therapy based on PET/CT staging and outperforms the traditional IPI in terms of discrimination and calibration. Additionally, an objective from a statistical point of view is to evaluate whether the more complex and elaborated modelling approaches are worth their effort and can help to better classify patients into groups of good and bad prognosis under (R-)CHOP therapy. Speaking of the three modelling approaches, logistic regression assumes that a prognostic variable contributes linearly to the odds of the event of interest, that is, a one-point increase in the prognostic variable goes along with the same increase in the odds irrespective of the variable's value. When modelling time-to-event endpoints using Cox regression, the linearity assumption applies to the natural logarithm of the relative hazard. Additionally, Cox regression assumes that the hazards of two groups defined by a prognostic variable are proportional over time (proportional hazards assumption). Numerous techniques have been proposed to

overcome violations of the proportional hazards assumption; the most simple by Cox himself in adding an interaction term of the given variable and the time variable to the model (Cox, 1972). The MFPT approach neither makes that assumption nor that of linearity. In their review and comparison work, Buchholz and Sauerbrei (2011) consider it among several other methods that release these two assumptions simultaneously. Thus, a third level of investigation in this thesis concentrates on the linearity and proportional hazard assumptions in time-to-event modelling, that is, the comparison between Cox regression and the MFPT approach. Taken together, this thesis shall develop a clinically meaningful and valid prognostic risk score model to improve decision making in aggressive NHL treatment, shall give guidance whether to use PET/CT staging in such a risk score, it shall provide support when performing categorisation of a given risk score, and shall also enrich the discussion on methodological aspects of prognostic model development and risk categorisation.

The remainder of this thesis reads as follows: The materials and methods section introduces the PETAL trial, describes the handling of the data, and explains the translation from prognostic models into risk scores and the measures that assess the scores' performance. The third section presents the results while the fourth section features their discussion. Subsequently, the thesis closes with a brief summary.

2 Material and methods

2.1 Data: The PETAL trial

The PETAL trial was a prospective multicenter randomised controlled study for patients with aggressive NHL initiated at the University Hospital Essen, University of Duisburg-Essen, Germany, by principal investigator Professor Dr. Ulrich Dührsen. Responsible biostatisticians were Professor Dr. Karl-Heinz-Jöckel during the planning phase, Professor Dr. André Scherag during the main part of the the study, and the author of this thesis during the analysis, reporting, and dissemination phase. The initial design of the trial was described by Dührsen et al. (2009). The main results were published by Dührsen et al. (2018) and a subgroup analysis of B-cell lymphoma patients was carried out by Hüttmann et al. (2019) while up-to-date work on several other manuscripts is in progress or already submitted for publication (Schmitz et al., 2019). The study was registered at EudraCT (European Union Drug Regulating Authorities Clinical Trials; EudraCT number 2006-001641-33) and at ClinicalTrials.gov (NCT00554164).

Fifty-seven oncological sites and twenty-three nuclear medicine institutions participated in the trial. The trial was conducted according to the guidelines of Good Clinical Practice and the principles stated in the latest revision of the Declaration of Helsinki. Also, ethics approval was obtained for every site and the study as a whole.

Patients with a positive baseline PET/CT received two cycles of (R-)CHOP which is the standard therapy for aggressive NHL. Based on the subsequent SUV-based interim PET/CT (iPET) interpretation, patients with favourable interim PET/CT were allocated to part A of the trial and patients with unfavourable interim PET/CT to part B. A favourable interim PET/CT response meant relative reduction of maximum SUV by more than 66 % compared to baseline (Lin et al., 2007). At the beginning of the study, patients in part A were uniformly treated with four additional cycles of (R-)CHOP. During the course of the study, newly published data (Pfreundschuh et al., 2008) motivated an amendment to the study protocol that introduced the randomisation of CD20-positive patients after two cycles of R-CHOP and interim PET/CT to either four additional cycles of R-CHOP or to the same treatment plus two extra doses of rituximab. After reaching the sample size goal for the randomised comparison regarding extra doses of rituximab, CD20-positive patients achieving a favourable interim PET/CT response were uniformly treated with four additional cycles of R-CHOP plus two extra doses of rituximab to further enrich this subgroup. All patients with an unfavourable interim PET/CT response (part B) were randomised to either continue R-CHOP for six additional cycles or receive six blocks of a more complex methotrexate-, cytarabine-, and etoposide-based regimen originally designed for Burkitt lymphoma (Hoelzer et al., 2014). As already mentioned in the introduction, rituximab was omitted in CD20-negative patients in both parts of the study. Those patients were uniformly treated with four and six additional cycles of CHOP in part A and B, respectively. For the sake of readability, this thesis refers to the treatment as (R-)CHOP irrespective of whether rituximab was part of the treatment regimen. The general design of the study with parts A and B and their respective treatment arms is shown in Figure 1. As described in this section, in contrast to patients in part B not all patients in part A underwent randomisation. Thus, this thesis will refer to the terms date of treatment allocation and interim PET/CT date instead of the more specific randomisation date. Sample size of the entire study population was based on the empirically derived assumption that treatment failure — a combined endpoint consisting of progression, relapse, treatment discontinuation due to toxicity, start of alternative therapy, and death of any cause — after two years could be improved from 80% to 90% in part A or from 30% to 45% in part B (alpha=0.05,

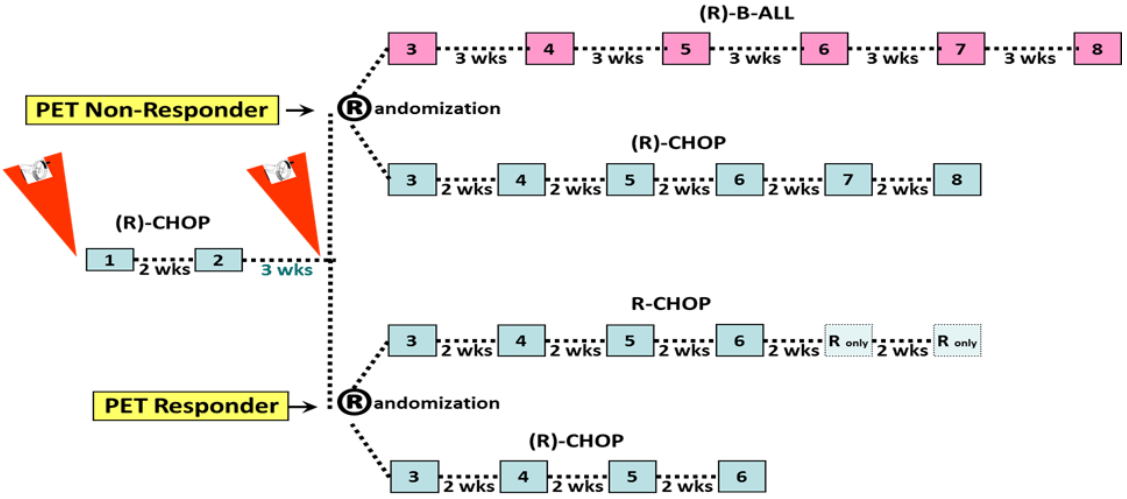


Figure 1: General design of the PETAL trial with part A ("PET responder") at the bottom and part B ("PET Non-Responder") at the top.

power=0.8; Dührsen et al., 2018).

Patient population and inclusion/exclusion criteria

Important inclusion criteria in the PETAL trial were newly diagnosed aggressive B-cell and T-cell NHL as proven histologically, age between 18 and 80 years, an Eastern Cooperative Oncology Group (ECOG) score (Oken et al., 1982) between zero and three, and a positive baseline PET/CT scan. Exclusion criteria included among others histological diagnosis of cerebral, Burkitt, and lymphoblastic lymphoma. Detailed inclusion and exclusion criteria were defined in the study protocol. Between 2007 and 2012, a total of 1072 patients were registered and 862 patients remained in the study until interim PET/CT to be allocated to a treatment arm — 754 in part A and 108 in part B. The group of allocated patients makes up the intention-to-treat population that this thesis focuses on. The CONSORT (Consolidated Statement of Reporting Trials) diagram (Schulz et al., 2010) of the PETAL trial depicts the flow of patients through the study and can be found in Figure 1 of Dührsen et al. (2018).

Endpoint and candidate variables For the purpose of this thesis, overall survival is used as endpoint to develop the prognostic risk scores — this decision is made despite the fact that the primary endpoint in the PETAL trial was time to treatment failure. Candidate variables for prognostic risk score models have to be well chosen and should be based on clinical knowledge from previous investigations. In consultation with the

clinical team, the set of candidate variables includes the IPI (Shipp et al., 1993) factors age (at treatment allocation after interim PET/CT), ECOG score (Oken et al., 1982), Ann Arbour staging (Carbone et al., 1971), number of extranodal manifestations, and lactate dehydrogenase (LDH). The ECOG score assesses the patient's performance status and daily living abilities. It ranges from zero to five where zero reflects an asymptomatic status and five refers to death. Note again that only patients with an ECOG score between zero and three were eligible for the PETAL trial. Ann Arbor staging characterises growth and spread of the lymphoma. It ranges from stage I to IV and indicates how many regions are affected by the tumour or whether there is extranodal involvement. While raw values of age are used and the variable is treated as continuous, LDH is considered as binary variable the same way as in the IPI indicating whether the respective value exceeds normal range or not. Regarding the ordinal variables ECOG score and Ann Arbour staging, it is checked by a likelihood ratio test whether they can be included in the risk score model as continuous variables or have to be treated as categorical variables. Furthermore, the plain number of extranodal manifestations as used in the IPI is replaced by an indicator variable for lymphomatous involvement in major organs (bone marrow, central nervous system, liver/gastrointestinal tract, or lung). Sex, histological diagnosis of the lymphoma, and B-lymphocyte antigen CD20 expression were stratification factors of the randomisation and are thus natural members of the candidate set. Note that histological diagnosis is divided into the four subgroups DLBCL, other B-cell lymphomas including PMBCL and FL (OthB), T-cell lymphomas (T), and other lymphomas (Other). The clinical team of the PETAL trial further identified variables that may also have an influence on overall survival and are candidates for the final risk score models as well. These are presence of B symptoms (fever, night sweats, and weight loss), obesity, and an indicator variable for completely resected manifestations. In this thesis, obesity is represented by body surface area (BSA). With weight and height corresponding to kilograms and centimetres, respectively, BSA is obtained by the Du Bois formula (Du Bois and Du Bois, 1916) as $BSA := 0.007184 \cdot weight^{0.425} \cdot height^{0.725}$. Finally, as PET/CT scanning was used to assess early treatment response in the PETAL trial, it enters the candidate set with the variables maximum SUV at baseline PET/CT scan and at interim PET/CT scan (Dührsen et al., 2018). Table 1 gives an overview of the abbreviations of the candidate variables that are used in the remainder of this thesis.

Table 1: Variables in the candidate set with their abbreviations and their respective level of measurement.

Variable	Abbreviation	Level of measurement
Age in years	Age	continuous
Eastern Cooperative Oncology Group index	ECOG	ordinal*
Ann Arbor staging	AnnArb	ordinal*
Lactate dehydrogenase above upper limit of normal range	LDH	binary
Lymphomatous involvement in major organs	MajOrg	binary
Sex	Sex	binary
Histological diagnosis of the lymphoma	Diag	categorical
CD20 expression	CD20	binary
B symptoms	BSympt	binary
Body surface area	BSA	continuous
Completely resected manifestations	Resect	binary
Maximum SUV at baseline PET/CT scan	SUVbase	continuous
Maximum SUV at interim PET/CT scan	SUVint	continuous

*Using a likelihood ratio test, ordinal variables are checked to be either included as categorical or continuous variables in the set of candidate variables.

2.2 Missing data analysis, data visualisation, and modelling assumptions

This subsection deals with the display of missing data as well as with the investigation of the underlying missing data mechanism. Missing data are a crucial issue in all statistical analyses as results can be biased when the missing data mechanism is not dealt with adequately. According to Little and Rubin (2002), missing data mechanisms of variables can be divided into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR means that the non-missing observations for a given variable are a random subset of all observations, that is, missingness of the variable is neither predicted by any variable in the data set nor by the missing value itself. When a variable is MAR, whether an observation of this variable is missing can be explained by other variables in the data set but is not dependent on the missing value itself. Finally, MNAR means that the missing value itself predicts the missingness of a variable, e.g., when higher values are more likely to be missing than lower values. For MCAR data, complete case analysis is an unbiased approach (Lachin, 1999) — however, associated with a loss in statistical power. For MAR, multiple imputation or maximum likelihood methods have to be used while MNAR data ask for even more elaborate methodology. In this thesis, complete case analysis is used and Section 4 discusses the validity of this approach from clinical and statistical point of view given the PETAL trial data. To underpin this discussion, the results section presents absolute and relative frequencies of missing data on variable, case, and observation basis, respectively, and displays the pattern of missing data graphically. The missingness mechanism is further explored using odds ratios from univariate logistic regression models with the missingness indicator as dependent variable and each candidate variable as independent variable. For potentially suspicious variables, possible reasons for missing data are discussed.

Follow-up information is illustrated graphically by reverse Kaplan-Meier curves (Schemper and Smith, 1996) and further investigated using the approach of Clark et al. (2002) termed completeness of follow-up. To describe the study population of the PETAL trial, variables from the risk score model candidate set are described using summary statistics appropriate for their respective level of measurement. That is, mean with standard deviation as well as median with first (Q1) and third quartile (Q3) for continuous variables and absolute and relative frequencies for binary and categorical variables. Additionally, the correlation between SUV at baseline and at interim PET/CT scan is assessed by Pearson's and Spearman's correlation coefficients. A likelihood ratio

test is used to verify whether ordinal variables should enter the set of candidate variables in continuous or categorical form. It tests the maximised log-likelihood of a Cox model utilising the respective variable in continuous form against the maximised log-likelihood of a Cox model treating the variable as categorical in comparing a transformation of their ratio with the critical value of a χ^2 distribution. If the test is statistically significant in favour of the model with the categorical variable, that is, the p-value is smaller than or equal to 0.05 and the log-likelihood is greater than that of the model with the continuous variable, the variable cannot be included in the candidate set as continuous variable. The outcome variable overall survival is illustrated as Kaplan-Meier curve (Kaplan and Meier, 1958). Furthermore, the number of events (deaths) and censored observations is given as well as median overall survival time and the cumulative 2-year overall survival probability.

To assess the necessity of using the MFPT approach as opposed to the Cox regression model, the linearity assumption is checked by investigating the functional forms of continuous variables with plots of cumulative martingale residuals across observed values. Additionally, plots of standardised cumulative Schoenfeld residuals over time shed light on indications for violations of the proportional hazards assumption. While martingale residuals (Barlow and Prentice, 1988) indicate whether a patient died earlier or later than expected given all of his or her clinical characteristics, Schoenfeld residuals (Schoenfeld, 1982) reveal whether the realisation of a variable was lower or higher than expected given the risk set of patients at the patient's time of death. Note that this means that Schoenfeld residuals are only defined for patients who suffer from the event of interest. For both residual plots, any systematic pattern different from the zero line suggests a lack of fit, i.e., hints for a violation of the respective assumption. An associated Kolmogorov-type supremum test based on $5,000$ patterns simulated from a scenario where the assumption holds then provides inference regarding violations of linearity and proportional hazards, see Lin et al. (1993) for details.

2.3 IPI-based and modelling-based prognostic risk scores

In this thesis five different risk scores are compared. Two are based on the IPI (Shipp et al., 1993) and three on the modelling approaches logistic regression (e.g., Hosmer et al., 2013), Cox regression (Cox, 1972), and the multivariable fractional polynomial time (MFPT) approach (Sauerbrei et al., 2007), respectively. This subsection describes the construction of these five risk scores.

IPI-based prognostic risk scores The IPI is widely used for risk prognosis in clinical practice and its original version (hereafter: IPI_{Shipp}) is used as reference for the other four risk scores. The IPI_{Shipp} consists of the factors age, ECOG score (Oken et al., 1982), Ann Arbour staging (Carbone et al., 1971) in its Cotswold modification (Lister et al., 1989), number of extranodal manifestations, and LDH. One point is added to the risk score when a factor meets a certain criterion. The higher the IPI_{Shipp} the higher the individual risk where age above 60 years, an ECOG score of at least 2, Ann Arbour stage III or IV, more than one extranodal manifestation (ExNod), and LDH above the upper limit of the normal range (ULN) are penalised with one point each. The resulting IPI_{Shipp} score then is the sum of penalty points. With χ_m, $m = 1, \ldots, 5$, being the realisations of indicator variables that contain the information on whether the criterion of the respective IPI_{Shipp} factor is fulfilled, the score reads as

$$\begin{aligned} IPI_{Shipp} = \sum_{m=1}^{5} \chi_m = \mathbf{I}_{(60,\infty)}(Age) + \mathbf{I}_{\{2,3\}}(ECOG) + \mathbf{I}_{\{III,IV\}}(AnnArb) \\ + \mathbf{I}_{(1,\infty)}(ExNod) + \mathbf{I}_{(ULN,\infty)}(LDH). \end{aligned}$$

It is then collapsed into groups of low risk (0–1 point(s)), low-intermediate risk (2 points), high-intermediate risk (3 points), and high risk (4–5 points).

With the IPI_{iPET}, this thesis proposes a natural extension of the IPI_{Shipp} that is based on the PETAL trial. It enhances the IPI_{Shipp} by the interim PET/CT result with an extra risk point for a ΔSUV below the cut-off value of 66% as specified in the PETAL study protocol and proposed by Lin et al. (2007). This extended IPI_{Shipp}-based risk score can be seen as simple predecessor of and motivation for the modelling-based risk scores in the next paragraphs. It is considered in the set of compared risk scores in this thesis to explore the added value of the interim PET/CT result beyond the classical IPI_{Shipp} factors. With χ_6 containing additional information on the ΔSUV variable, it reads as

$$IPI_{iPET} = \sum_{m=1}^{6} \chi_m = IPI_{Shipp} + \mathbf{I}_{(-\infty,66\%)}(\Delta SUV).$$

To avoid loss of information arising from a categorisation into arbitrary classes from low risk to high risk levels, the IPI_{iPET} results are not collapsed but remain in the range $[0, 6]$.

Modelling-based prognostic risk scores Binary logistic regression, Cox proportional hazards regression, and the MFPT approach are used in this thesis to respectively construct modelling-based prognostic risk scores. Note that regarding the modelling approaches the term "risk score" is rather liberal use as the methods either model the odds of the event at a specific time point (logistic regression) or the hazard rate over time (Cox regression and MFPT approach) instead of the risk at a certain time point itself. The general setup for the three models is one and the same: Let there be n observations each with an outcome variable and a candidate set of potential influential factors and covariates — represented by the random variables $X_1, \ldots, X_K$. Logistic regression, Cox regression, and MFPT approach have the same aim to explain the relationship between the prognostic variables $X_1, \ldots, X_K$ and the outcome variable. Backward elimination is used as variable selection procedure for all three methods where for all variables the significance level for elimination is set to 0.01 to include strong predictors only (Sauerbrei et al., 2007). Interaction terms are not considered in the risk score models since the multivariable fractional polynomial regression does not yet allow for it when using time-to-event data with time-dependent effects, that is, using the MFPT approach. All three approaches provide the possibility to define a general prognostic risk score in the form

$$R_{general} := \sum_{k=1}^{K} x_k \omega_k$$

where x_k, $k = 1, \ldots, K$, are the realisations of the variables in the final model (selected by backward elimination) and ω_k are the variables' weights, that is, the estimated regression coefficients obtained by the respective modelling approach. The remainder of this subsection explains how logistic regression, Cox regression, and MFPT approach differ in their general model formulation, and thus, in the way the weights ω_k are estimated. To be notationally consistent across all three modelling approaches, the notation of Royston and Sauerbrei (2008) serves as an orientation point. The similarity of the resulting modelling-based risk scores is assessed in Section 3.2 that provides scatter plots and Pearson's correlation coefficients between the three scores.

Binary logistic regression In logistic regression (LogReg; for details see, e.g., Hosmer et al., 2013) the outcome is a binary random variable Y that follows a Bernoulli distribution while $X_1, \ldots, X_K$ are prognostic variables as described before. That is,

the event of interest occurs with probability π — or in other words $\pi(\boldsymbol{x}) = E(Y|\boldsymbol{x})$ where $\boldsymbol{x} = (x_1, \ldots, x_k)'$ is the vector of the realisations of $X_1, \ldots, X_K$. The actual endpoint in this thesis is overall survival which is a time-to-event instead of a binary variable. The time dimension from this endpoint is ignored for the logistic regression analysis and the binary variable death within two years from interim PET/CT is used. A justification for the choice of the 2-year time point can be found in the discussion. Logistic regression models the logarithm of the odds or chance $\pi/(1-\pi)$ of the event Y given the realisations of the random variables $X_1, \ldots, X_K$. The term $\log\left(\frac{\pi}{1-\pi}\right)$ is known as the logit of π. Therefore logistic regression is also known as the logit model that is defined as

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{k=1}^{K} x_k \beta_k.$$

Analogous to linear regression, the interpretation of the regression coefficients β_k is that a one-point increase in x_k leads to an increase of β_k in $\text{logit}(\pi)$. When X_k is a categorical variable, this applies to the increase in $\text{logit}(\pi)$ with respect to a pre-defined reference category of X_k. The term $\exp(\beta_k)$ is known as the odds ratio, i.e., the quotient between two groups regarding the chance of experiencing the event. Note that groups can again either be defined by a categorical variable (e.g., sex with male versus female) or by a continuous variable (e.g., age). In the latter case, the chance of suffering the event increases by the factor $\exp(\beta_k)$ when the continuous variable is increased by one unit. If the rare disease assumption is fulfilled, the odds ratio can serve as reliable estimate for the relative risk — for example in case-control studies where the prevalence of the investigated disease is low and the relative risk cannot be determined. Note that odds ratios are not used for the construction of the risk scores but in other parts of this thesis when logistic regression is applied. Instead, in using ordinary least squares methods to obtain the maximum likelihood estimators $\widehat{\beta}_k$, $k = 1, \ldots, K$, of the regression coefficients, the weights ω_k in the formulation of $R_{general}$ are replaced to construct a prognostic risk score R_{LogReg} for logistic regression as

$$R_{LogReg} = \sum_{k=1}^{K} x_k \widehat{\beta}_k.$$

Cox proportional hazards regression The Cox regression model (CoxReg; Cox, 1972) incorporates the time until the event of interest occurs, i.e., it can handle right-censored survival data as it is the case for overall survival. Let Z be a random variable

for survival time representing the time between an initial time point $t_0 = 0$ (here: interim PET/CT) and the occurrence of the event of interest (here: death). Let further be C a random variable for censoring time where Z and C are assumed to be independent. Then, T serves as outcome variable in Cox regression, and for patient n, $n = 1, \ldots, N$, $t_n = \min(z_n, c_n)$ is the observed time since interim PET/CT (t_0) with δ_n indicating whether that time interval ends with the patient's death ($\delta_n = 1$; $t_n = z_n$) or not ($\delta_n = 0$; $t_n = c_n$). Given t_0 and the vector $\boldsymbol{x}$, Cox regression models the time-dependent hazard λ of an event as

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp\left(\sum_{k=1}^{K} x_k \gamma_k\right).$$

In comparison to the logit model, the intercept β_0 is replaced by the time-dependent baseline hazard function λ_0 in the Cox model. Analogously to the odds ratio in logistic regression, the term $\exp(\gamma_k)$ is known as the hazard ratio (HR) between two groups defined by the variable X_k. Note again that hazard ratios are not used for the construction of the risk scores but only in other parts of this thesis. Treating λ_0 as an unspecified increasing function, the parameter vector γ_k, $k = 1, \ldots, K$, can be estimated via partial likelihood to define a prognostic risk score R_{CoxReg} for the Cox model as

$$R_{CoxReg} = \sum_{k=1}^{K} x_k \widehat{\gamma}_k.$$

Multivariable fractional polynomial time approach As mentioned in the introduction, logistic regression assumes linearity between continuous covariates and the logit while Cox regression assumes linearity between continuous covariates and the natural logarithm of the relative hazard, i.e., $\ln\left(\frac{\lambda(t)}{\lambda_0(t)}\right)$. Also, Cox regression assumes proportionality of the group hazards over time where groups can be defined by a categorical as well as by a continuous variable. The multivariable fractional polynomial (MFP) procedure by Sauerbrei and Royston (1999) accounts for violations from the linearity assumption in the class of generalised linear models that, for example, includes linear and logistic regression. It allows to employ other functional forms than the linear relationship for continuous variables while the MFP time (MFPT) approach (Sauerbrei et al., 2007; Royston and Sauerbrei, 2008, Chapter 11) extends this approach to right-censored data and additionally accounts for the possibility of non-proportional hazards in time-to-event modelling. While non-linearity is facilitated as in any other

generalised MFP model in using transformations of X_k, the latter is achieved by allowing the regression coefficient to vary in time, i.e., being a function of t — where Sauerbrei et al. (2007) suggest to model the time-dependency of a variable X_k using a fractional polynomial function φ (Royston and Altman, 1994). Note that the relaxation of the linearity assumption means that the variable X_k may already enter this modelling of time-dependency as a non-linear transformation, that is, also being modelled as a fractional polynomial function. In the univariate case, the approach of Sauerbrei et al. (2007) is termed the fractional polynomial time (FP-time) procedure. Although this procedure can be generalised to a fractional polynomial function of arbitrary degree L, $L > 0$, this thesis only makes use of fractional polynomials with a maximum degree of $L = 2$. In this sense, a fractional polynomial function of first degree ($L = 1$) is defined as

$$\varphi_1(t, q) = \psi_0 + \psi_1 t^q$$

whereas a fractional polynomial function of second degree ($L = 2$) is defined as

$$\varphi_2(t, (q_1, q_2)) = \psi_0 + \psi_1 t^{q_1} + \psi_2 t^{q_2} = \begin{cases} \psi_0 + \psi_1 t^{q_1} + \psi_2 t^{q_2}, & q_1 \neq q_2 \\ \psi_0 + \psi_1 t^{q_1} + \psi_2 t^{q_2} \log(t), & q_1 = q_2 \end{cases}.$$

Here, q and (q_1, q_2) are the powers of the respective fractional polynomial function coming from the standard set of powers $\mathcal{Q} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, and ψ_0, ψ_1, and ψ_2 are the parameters of the fractional polynomial function. By convention, $t^0 = \log(t)$ and note that $t^0 = \log(t + 1)$ is used when there are patients with $t = 0$. Figure 2 presents a selection of possible fractional polynomial transformations of first and second degree where the powers q, q_1, and q_2 determine the general shape of the transformation and ψ_0, ψ_1, and ψ_2 the magnitude of the shape. As can be seen from Figure 2, first-degree fractional polynomials represent well-known transformations like, e.g., quadratically increasing and decreasing functions as well as other variations of convex and concave functions. Second-degree transformations are, in contrast, more flexible and also include more complex shapes that may even differ considerably within a certain class of fractional polynomials with the same powers. Already this brief selection of possible transformations illustrates that the eight first-degree and thirty-six second-degree power combinations for fractional polynomial functions cover a broad number of non-linear transformations.

For a second-degree fractional polynomial transformation of t, the non-proportional

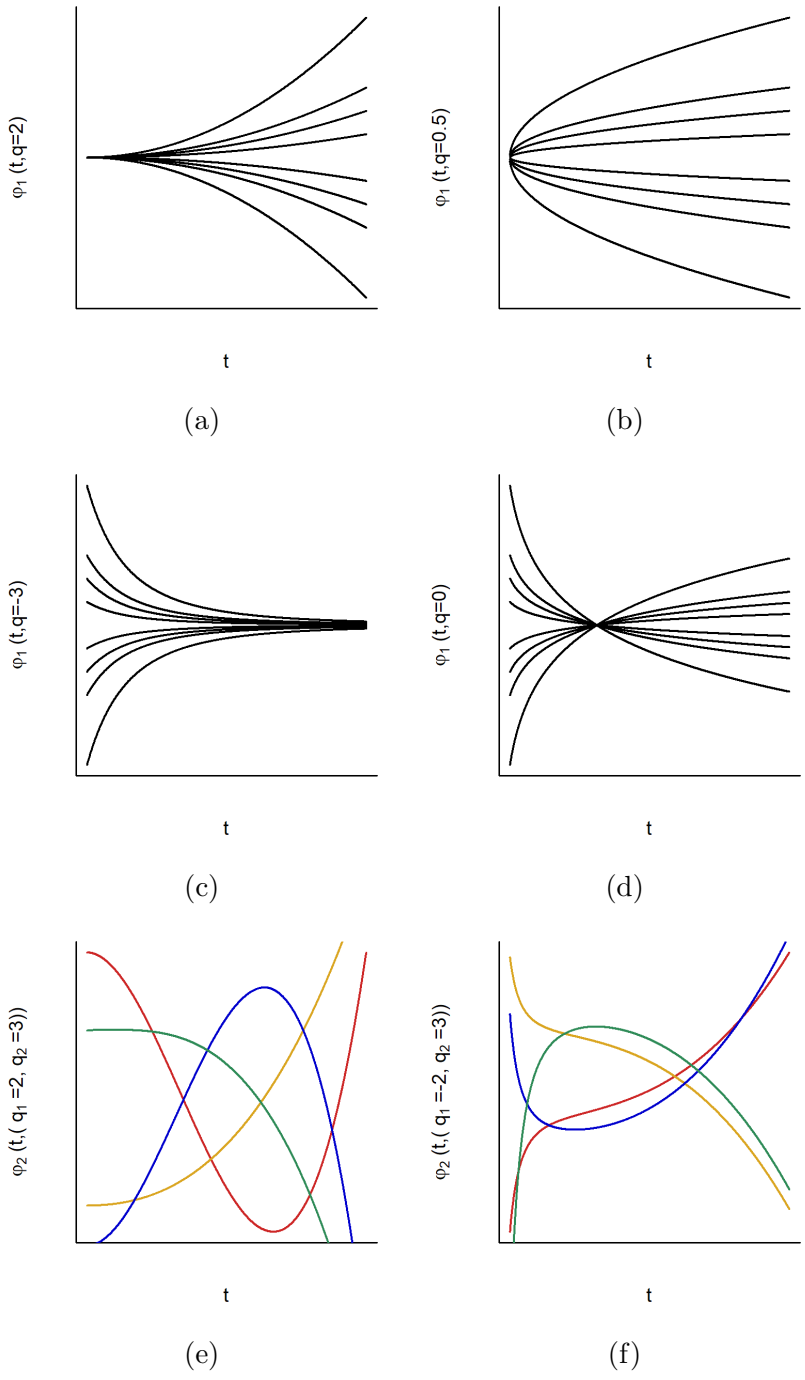


Figure 2: Possible shapes of first-degree fractional polynomial functions with power (a) $q = 2$, (b) $q = 0.5$, (c) $q = -3$, and (d) $q = 0$ as well as of second-degree fractional polynomial functions with powers (e) $q_1 = 2$ and $q_2 = 3$ and (f) $q_1 = -2$ and $q_2 = 3$. The final appearance depends in addition to the selected power(s) also on the estimated regression coefficients ψ_0, ψ_1, and ψ_2.

hazards model from the FP-time procedure for a variable X_k becomes

$$\begin{aligned} h(t|x_k) &= h_0(t)\exp\left\{x_k\varphi_2(t,(q_1,q_2))\right\} \\ &= \begin{cases} h_0(t)\exp\left\{x_k(\psi_0+\psi_1 t^{q_1}+\psi_2 t^{q_2})\right\}, & q_1 \neq q_2 \\ h_0(t)\exp\left\{x_k(\psi_0+\psi_1 t^{q_1}+\psi_2 t^{q_2}\log(t))\right\}, & q_1 = q_2 \end{cases} \end{aligned}$$

while for degree $L = 1$ it analogously reads as

$$h(t|x_k) = h_0(t)\exp\left\{x_k\varphi_1(t,q)\right\} = h_0(t)\exp\left\{x_k(\psi_0+\psi_1 t^{q})\right\}.$$

In their textbook, Royston and Sauerbrei (2008, Chapter 4.10.2) describe the way of eliciting the optimal fractional polynomial function $\varphi^*(t,\boldsymbol{q})$ for a variable X_k as function selection procedure where $\boldsymbol{q}$ denotes the vector of powers of the optimal fractional polynomial function. Note that the dimension of $\boldsymbol{q}$ is unknown beforehand and depends on the degree of the optimal fractional polynomial function. The function selection procedure is based on the closed testing principle (Marcus et al., 1976) and consists of three steps. These steps compare the deviances of the best fractional polynomial models of first and second degree against each other and against the deviances of a null and a default model. For this purpose, deviances of models are defined as minus twice the maximised partial likelihood for $h(t|x_k)$. The best model of a given degree L is denoted as FP_L^* and defined as the model that minimises the deviance among the models of the same degree. It is elicited through a grid search on all possible power combinations of the given degree in the standard power set $\mathcal{Q}$. The first step of the function selection procedure tests whether there is an overall association between the outcome and the most complex interaction of X_k and time, that is, whether the variable needs to be modelled time-dependently at all. Thus, the best second-degree fractional polynomial model is tested at significance level α_0 against the null model that assumes a time-constant effect. Note that for a variable that has to undergo the complete function selection procedure due to, e.g., prior knowledge or design reasons, $\alpha_0 = 1$. If the test is significantly in favour of FP_2^*, it will be tested in the second step at significance level α_1 against the default model with $\varphi_1(t, q = 0) = \log(t)$ that represents the most common functional form in survival data. If that test also favours FP_2^*, the last step will test it at significance level α_2 against the best first-degree model — FP_1^*. The procedure ends when FP_2^* has been tested against FP_1^* or has not been superior to the null or the default model. The optimal polynomial function $\varphi^*(t,\boldsymbol{q})$ then either is the

one associated with FP_2^* or the one associated with the model that FP_2^* failed to be superior to.

The MFPT approach extends the FP-time procedure with its function selection procedure to the case of K variables. The algorithm consists of three stages: In the first stage, the MFP approach (Sauerbrei and Royston, 1999) is used to build a model $\mathcal{M}_0$ including time-fixed effects only. The MFP approach starts with the full model with all candidate variables included and allows for dropping uninfluential variables from the model in using backward elimination ($\alpha = 0.01$). Fractional polynomial functions for the continuous variables are determined by the function selection procedure for each variable separately while adjusting for all other variables currently in the model. The final model $\mathcal{M}_0$ is obtained when selected variables and fractional polynomial functions have converged. The second stage of the MFPT algorithm investigates whether any variable not included in $\mathcal{M}_0$ only has a short-term effect on the outcome instead of an effect over the whole time period. Therefore, the MFP approach is repeated on a data set where all events after a time point t_{short} are censored while adjusting for all variables included in $\mathcal{M}_0$ and with their regression coefficients (but not their powers) being re-estimated. Variables with a statistically significant short-term effect are added to $\mathcal{M}_0$ to obtain $\mathcal{M}_1$. Note that for the time point t_{short}, also two years after interim PET/CT is used. In the third stage, time-varying effects of the variable are detected and modelled in subsequently applying the FP-time procedure to each variable in $\mathcal{M}_1$ while adjusting for all other variables in the model (Sauerbrei et al., 2007). In this thesis, the significance level with respect to time-dependency of a variable's effect is set to 0.05. The vector of the realisations of the variables $\boldsymbol{x} = (x_1, \ldots, x_K)'$ and the vector of their respective optimal fractional polynomial transformations $\boldsymbol{\varphi}^*(t, (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K)) = (\varphi_1^*(t, \boldsymbol{q}_1), \ldots, \varphi_K^*(t, \boldsymbol{q}_K))'$ then make up the model $\mathcal{M}_2$. Here, $\boldsymbol{q}_k$, $k = 1, \ldots, K$, are the vectors of the powers of the optimal fractional polynomial functions for $\boldsymbol{x}$. As mentioned above, the dimension $\dim(q_k)$ depends on the selected degree of the optimal fractional polynomial function. That is, $\dim(q_k) = 2$ for an optimal fractional polynomial function of degree two, $\dim(q_k) = 1$ for degree one, and $\dim(q_k) = 0$ for a time-constant variable with no fractional polynomial transformation being selected. The MFPT non-proportional hazards model then reads as

$$h(t|\boldsymbol{x}) = h_0(t) \exp\left\{\boldsymbol{x}'\boldsymbol{\varphi}^*(t, (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K))\right\}$$

where the linear combination of the fractional polynomial transformations can be written

as

$$\begin{aligned}\boldsymbol{x}'\boldsymbol{\varphi}^*(t,(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K)) = &\sum_{x_k\in\mathcal{M}_2;\,\dim(q_k)=2} x_k(\psi_{k,0}+\psi_{k,1}t^{q_{k,1}}+\psi_{k,2}t^{q_{k,2}})\\ &+\sum_{x_k\in\mathcal{M}_2;\,\dim(q_k)=1} x_k(\psi_{k,0}+\psi_{k,1}t^{q_k})\\ &+\sum_{x_k\in\mathcal{M}_2;\,\dim(q_k)=0} x_k\psi_{k,0},\end{aligned}$$

$$t^{q_{k,2}} = \begin{cases} t^{q_{k,2}}, & q_{k,2}\neq q_{k,1}\\ t^{q_{k,2}}\log(t), & q_{k,2}=q_{k,1}\end{cases}.$$

The first sum is over the variables with second-degree fractional polynomial transformations, the second sum over those with degree one (including the default model), and the third sum is over all time-constant variables. Note that in the latter case $\psi_{k,0}$ will be equivalent to the coefficient γ_k from Cox regression if the function selection procedure does not identify a time-dependent effect for any variable. Also note that still, $t^0 = \log(t)$ by definition. The linear combination of the optimal fractional polynomial functions from above can with its estimated powers and parameters then be used analogously to logistic and Cox regression to construct a risk score for a given person as

$$R_{MFPT} = \boldsymbol{x}'\widehat{\boldsymbol{\varphi}}^*(t,(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K)).$$

That is, the weights ω_k from $R_{general}$ are fractional polynomial functions of varying degree that are selected and estimated by the function selection procedure.

When calculating the risk score R_{MFPT} for a new patient with given clinical data and tumour characteristics, the question arises how to incorporate a time-dependent function into the risk score as at the time of diagnosis both the future observation and the survival time of the new patient will always be unknown. The approach used in this thesis is rather simple and represents a weighted mean of the fractional polynomial function with the weights being the relative frequencies of the observed censoring and survival times in the entire data set. Thus, with n_t being the number of patients terminating the study at observation time t and with t_{max} being the maximum observation time in the sample,

the weight ω_k for a variable X_k with a time-dependent coefficient can be estimated as

$$\widehat{\omega}_k = \frac{1}{N} \sum_{t=0}^{t_{max}} n_t \cdot \widehat{\varphi}_k^*(T = t, \boldsymbol{q}_k)$$

where $N = \sum_{t=0}^{t_{max}} n_t$ is the total sample size and $\widehat{\varphi}_k^*(T = t, \boldsymbol{q}_k)$ the evaluation of the selected fractional polynomial function at time point t. More sophisticated alternatives to this weighting approach that consider the time-dependency in different ways are discussed in Section 4.

2.4 Assessment of discrimination and calibration

Among the many different kinds of measures for predictive accuracy of prognostic risk models, discrimination has the most straightforward interpretation from a clinical point of view (Antolini et al., 2005). It evaluates whether patients with higher risk score values as elicited from a prognostic model also have worse outcomes than patients with lower risk score values. Diagnostic performance measures from this class can be divided into global and local measures where a global measure rates the general ability of the discrimination variable (over the whole range of that variable) to separate patients with good and bad outcomes, respectively. A local discrimination measure assesses the separation of good and bad outcome patients after defining a binary classification rule based on a specific cut-off or operating point. Risk scores from prognostic models can then be used to make treatment decisions depending on the individual risk for each patient. The resulting cut-off point for the continuous risk score reflects the decision for a — say more aggressive but also more toxic — therapy when that certain value of the risk score is exceeded. Note that each value of the risk score, and thus, each possible cut-off point also corresponds to a certain risk of experiencing the event.

Discrimination In this thesis, a score's ability to discriminate patients with good outcome from those with bad outcome is assessed on the basis of overall survival — irrespective of whether the risk score is elicited from a binary or a time-to-event endpoint. Thus, for the definition of the classifier performance measures utilised here, the basic classification measures sensitivity and specificity have to be introduced in terms of right-censored time-to-event data (Rota et al., 2015). Assume that a time point of interest τ is given and also a cut-off point ρ that divides the realisations of a risk score variable R into testing positive and negative. Then, a patient is said to have experienced

the event up to time point τ when for the realisation of the random variable Z holds $z_n \leq \tau$, and is said to be event-free when $z_n > \tau$. Consequently, sensitivity (SE) is a function of ρ and defined as the probability of a positive test given the patient has died before or at time point τ:

$$SE(\rho) = \Pr(R > \rho | Z_n \leq \tau).$$

Likewise, specificity (SP) is defined as the probability of a negative test given the patient is still alive at τ:

$$SP(\rho) = \Pr(R \leq \rho | Z_n > \tau).$$

The main issue in calculating these measures for right-censored event-time data is to incorporate the information of the patients whose survival times are censored before τ. According to Antolini and Valsecchi (2012), the initial classification matrix would appear as depicted in Table 2. In Section 3.1, the distribution of the censoring status with respect to the time point τ is given by the realisations of N_D, $N_{\bar{D}}$, and N_C from Table 2, that is, the numbers of patients who died before or at τ, of patients who are known to have survived at least until τ, and of patients whose censoring time is shorter than τ.

Table 2: Initial classification matrix according to Antolini and Valsecchi (2012). D is for deceased until τ, $\bar{D}$ for alive at τ, and C for censored before τ.

	Deceased until τ	Alive at τ	Censored before τ	Total
Testing positive ($R > \rho$)	N_D^{pos}	$N_{\bar{D}}^{pos}$	N_C^{pos}	N^{pos}
Testing negative ($R \leq \rho$)	N_D^{neg}	$N_{\bar{D}}^{neg}$	N_C^{neg}	N^{neg}
Total	N_D	$N_{\bar{D}}$	N_C	N

To obtain a classical 2×2 classification matrix as known from discrimination measures for binary classifiers, Antolini and Valsecchi (2012) introduce an approach that makes use of the data from the censored observations before τ to enrich the dead and alive at τ numbers from Table 2. The most promising techniques they propose are direct estimation by conditional weighting or by imputation which they show to be equivalent. In this thesis, the latter approach is used as it is notationally easier and more intuitive because it utilises standard time-to-event methods. With the imputation approach, the N_C censored observations are assigned a probability that they are deceased or not

depending on their observed censoring time. The procedure has to be performed for the population of the testing positive and the testing negative patients separately to obtain probabilities for the N_C^{pos} and N_C^{neg} censored observations, respectively. Without loss of generality, a censored observation from the testing positive population contributes to the non-deceased part with probability

$$\Pr(Z > \tau | Z > T_j; \mathbf{I}_{(\rho,\infty)}(R)) = w_j,\ j = 1, \ldots, N_C^{pos},$$

and to the part of deceased patients with probability

$$\Pr(Z \leq \tau | Z > T_j; \mathbf{I}_{(\rho,\infty)}(R)) = 1 - w_j.$$

The weights w_j for the testing positive population are estimated as the fraction of the Kaplan-Meier estimates (Kaplan and Meier, 1958) $\hat{S}$ at τ and at T_j, and thus, as $\hat{w}_j = \hat{S}^{pos}(\tau)/\hat{S}^{pos}(T_j)$. The resulting enriched 2×2 classification matrix takes a form as shown in Table 3 where the estimated weights for the testing negative population are analogously defined and denoted as $\hat{w}_i$, $i = 1, \ldots, N_C^{neg}$ (for more details, see Antolini and Valsecchi, 2012).

Table 3: Enhanced 2×2 classification matrix as proposed by Antolini and Valsecchi (2012). D is for deceased until τ, $\bar{D}$ for alive at τ.

	Deceased until τ	Alive at τ	Total
Testing positive ($R > \rho$)	$N_D^{pos} + \sum_{j=1}^{N_C^{pos}} (1 - \hat{w}_j)$	$N_{\bar{D}}^{pos} + \sum_{j=1}^{N_C^{pos}} \hat{w}_j$	N^{pos}
Testing negative ($R \leq \rho$)	$N_D^{neg} + \sum_{i=1}^{N_C^{neg}} (1 - \hat{w}_i)$	$N_{\bar{D}}^{neg} + \sum_{i=1}^{N_C^{neg}} \hat{w}_i$	N^{neg}
Total	$N_D + \sum_{j=1}^{N_C^{pos}} (1 - \hat{w}_j) + \sum_{i=1}^{N_C^{neg}} (1 - \hat{w}_i)$	$N_{\bar{D}} + \sum_{j=1}^{N_C^{pos}} \hat{w}_j + \sum_{i=1}^{N_C^{neg}} \hat{w}_i$	N

Sensitivity and specificity can be estimated from that 2×2 classification matrix the same way as in the binary case with $\tilde{N}_D := N_D + \sum_{j=1}^{N_C^{pos}} (1 - \hat{w}_j) + \sum_{i=1}^{N_C^{neg}} (1 - \hat{w}_i)$ and $\tilde{N}_{\bar{D}} := N_{\bar{D}} + \sum_{j=1}^{N_C^{pos}} \hat{w}_j + \sum_{i=1}^{N_C^{neg}} \hat{w}_i$ as

$$\widehat{SE}(\rho) = \frac{N_D^{pos} + \sum_{j=1}^{N_C^{pos}} (1 - \hat{w}_j)}{\tilde{N}_D}$$

and

$$\widehat{SP}(\rho) = \frac{N_{\bar{D}}^{neg} + \sum_{i=1}^{N_C^{neg}} \hat{w}_i}{\tilde{N}_{\bar{D}}}$$

and can be represented graphically as receiver operating characteristic (ROC) curve plotting $\widehat{SE}$ across $1 - \widehat{SP}$. Note that in constructing a 2×2 classification matrix as described before, the approach of Antolini and Valsecchi (2012) allows for the use of standard ROC analysis instead of methodology specifically designed for time-to-event endpoints. And in using standard ROC methods, each point on the ROC curve represents a possible cut-off point that is associated with specific sensitivity and specificity. Plotting the ROC curve also illustrates the graphical interpretation of the three most commonly used measures of local discrimination performance (see Figure 3): Youden index (Youden, 1950), concordance probability (Liu, 2012), and closest-to-(0,1) corner criterion (Perkins and Schisterman, 2006). These measures are regularly used in practice to inform cut-off point selection in determining the optimal operating point with respect to the given criterion. In this thesis, the primary diagnostic performance measure is the closest-to-(0,1) corner criterion. It is defined as the Euclidian distance (ED) between the ROC curve and the upper left corner in the ROC curve plot:

$$ED(\rho) = \sqrt{(1 - SE(\rho))^2 + (1 - SP(\rho))^2}.$$

To be robust against the choice of the diagnostic performance criterion, the cut-off points obtained by Youden index (also called J index; J) and concordance probability (CP) are reported as sensitivity analyses for all obtained risk scores. The Youden index is defined as

$$J(\rho) = SE(\rho) + SP(\rho) - 1$$

and can be interpreted graphically as the vertical distance between the ROC curve and the diagonal chance line. The graphical interpretation of the concordance probability is a rectangle of width SP and height SE, and thus, calculated as

$$CP(\rho) = SE(\rho) \cdot SP(\rho).$$

While the closest-to-(0,1) corner method ED is a minimisation problem in ρ, optimal cut-off points using Youden index and concordance probability are found in maximising J and CP, respectively. In this thesis, the determination of the optimal cut-off point using these three approaches is restricted to the interval between the 5th and the

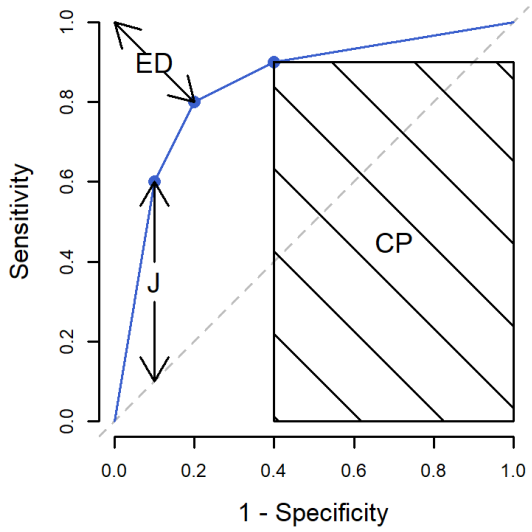


Figure 3: Graphical explanation of the three local discrimination measures closest-to-(0,1) corner criterion (ED), Youden index (J), and concordance probability (CP) — illustrated at three different possible points of a hypothetical ROC curve (in blue).

95th quantile of the respective risk score variable R to exclude extreme cut-off points that would result in almost all patients being tested either positive or negative. The results section reports the cut-off points for all five risk scores obtained by using the three selection criteria ED, J, and CP. Furthermore, the resulting prognostic groups according to the primary criterion ED are compared on a score-by-score basis and the concordance of the prognosis classification of two risk scores is assessed by their fraction of agreement.

Additionally, as any proposed risk prediction model is expected to be accompanied by a global discrimination measure (Pencina et al., 2010), the area under the ROC curve (AUC) is given. Let ρ_u, $u = 1, \ldots, U$, be U cut-off points that divide the risk score R in two non-empty parts with given sensitivity and specificity corresponding to the points on the ROC curve. Note that U corresponds to the number of unique realisations of R minus one. By convention, let $SE(\rho_0) = 1$ and $SP(\rho_0) = 0$ as well as $SE(\rho_{U+1}) = 0$ and $SP(\rho_{U+1}) = 1$. The area under the ROC curve can then be estimated with the trapezoidal rule as

$$\widehat{AUC} = \sum_{u=1}^{U+1} \frac{SE(\rho_u) + SE(\rho_{u-1})}{2} (SP(\rho_u) - SP(\rho_{u-1}))$$

given the sensitivity–specificity pairs of the cutoff points ρ_u. For an optimal classifier, the AUC equals one while for the worst classifier $AUC = 0.5$. In the latter case, the ROC curve coincides with the diagonal chance line from $(0, 0)$ to $(1, 1)$, and in the optimal case, it goes from $(0, 0)$ to $(0, 1)$ to $(1, 1)$. According to Hanley and McNeil (1982), with $A_1 = AUC/(2 - AUC)$ and $A_2 = 2 \cdot AUC^2/(1 + AUC)$ an approximate standard error of the area under the ROC curve is given as

$$\widehat{\sigma}_{AUC} = \sqrt{\frac{AUC(1 - AUC) + (\tilde{N}_D - 1)(A_1 - AUC^2) + (\tilde{N}_{\bar{D}} - 1)(A_2 - AUC^2)}{\tilde{N}_D \cdot \tilde{N}_{\bar{D}}}}$$

and can be used to construct a 95% confidence interval (CI) for the area under the ROC curve that reads as

$$CI_{AUC} = AUC \pm 1.96 \cdot \sigma_{AUC}.$$

Calibration Alongside discrimination, calibration is an important tool to assess the performance of a prognostic model. It is a measure of predictive accuracy, that is, observed probabilities are compared with the predicted probabilities elicited from the given prognostic model. In this thesis, the Brier score (Brier, 1950) is used to assess predictive accuracy. Initially introduced for binary and categorical outcomes in meteorology, Graf et al. (1999) adapted the Brier score to time-to-event outcomes. In this adaptation, it is applied to a Cox model for overall survival with the risk score under investigation being used as independent variable. Consequently, it follows that the risk scores are assumed to contribute linearly to the natural logarithm of the relative hazard and that also the proportional hazards assumption holds. "[The Brier score] is given by the squared distance between the patients' observed status and the predicted probability. The values of the Brier score can be interpreted as the loss or regret which is incurred when the prediction is issued to a patient whose true status is the observed status." (Gerds et al., 2008). Note that this means the Brier score (BS) can be interpreted as prediction error, and thus, the lower the score the better the calibration performance. In survival analysis, it is defined time-dependently as weighted mean squared error of prediction as

$$\widehat{BS}(t, S) = \frac{1}{N} \sum_{n=1}^{N} \widehat{\eta}_n(t)(y_n(t) - \widehat{S}_n(t, \boldsymbol{x}_n))^2$$

where $y_n(t)$ is the observed survival status at time point t and $\widehat{S}_n(t, \boldsymbol{x}_n)$ the respective predicted survival probability of patient i at time point t estimated by the Cox model.

The weights $\widehat{\eta}_n(t)$ account for right-censoring and are obtained by inverse-probability censoring weighting based on the marginal Kaplan-Meier estimator, that is, ignoring any predictor variables (Mogensen et al., 2012). When the Brier score is followed over time, it can be displayed as prediction error curve. Then, the integrated Brier score (IBS) serves as summarising measure for the risk score's calibration performance and is elicited from the area under the prediction error curve as

$$\widehat{IBS}(BS, t_0, t_{max}) = \frac{1}{t_{max}} \int_{t_0}^{t_{max}} BS(u, S) \mathrm{d}u$$

where t_{max} is the maximum of observed censoring and event times for which a prediction error can be obtained.

2.5 Validation strategies

Validation can be used to verify the performance of a prognostic model in terms of discrimination and calibration. Particularly, validation offers an opportunity to evaluate "optimism of the result" and to "over insights into the 'true' model reproducibility" (Antolini et al., 2005). Validation strategies can be divided into three categories (Altman and Royston, 2000): Internal validation where resampling techniques are used to test the prognostic risk score on the same data set as it was developed on, temporal validation where the test data set contains patients from the same institutions as the development data set but with temporal separation between the two, and external validation with the test data set including completely independent patients from different institutions than the training sample. Due to the lack of temporally separated or completely independent data, only internal validation on the PETAL data set (cf. Section 2.1) can be applied in this thesis. For the internal validation of the five risk scores, an 8-fold cross-validation (Stone, 1974; Geisser, 1975) is used where the data set is divided into eight sub-data sets of equal size. In the cross-validation process with eight repetitions, each sub-data set serves as test data set once while the remaining seven parts serve as training data set. For each repetition, a prognostic model is developed on the training data set and the resulting model is then validated on the new or unknown data of the test data set. To avoid each of the eight validation samples being a random subset of the entire data, the data set is separated according to the patients' treatment allocation dates — artificially creating a time-separated structure (Altman and Royston, 2000). Note, however, that this approach is not equal to a temporal validation as training and test data patients

from one and the same study will most likely not be independent from each other. The time-separated structure of the validation samples is displayed graphically, and for all performance measures results are presented by validation sample.

3 Results

This section presents the results of the thesis that were obtained using the statistical software package R (Version 3.5.1; R Core Team, Vienna, Austria), SAS software (Version 9.4; SAS Institute Inc., Cary, NC), and — for the MFPT approach — Stata statistical software (Version 11; StataCorp LP, College Station, TX).

3.1 Missing data analysis, data visualisation, and modelling assumptions

The PETAL data set contains only very few missing data. There are 33 missing values in nine of the thirteen candidate variables that are distributed across 15 of the 862 patients. Likewise, there is no conspicuous pattern in missing data (Figure 4).

Univariate logistic regression for a missing data indicator variable (Table 8 in the Appendix) reveals slightly higher odds of having missing data for patients with higher maximum SUV at interim PET/CT. However, this finding is based on only ten patients with available maximum SUV at interim PET/CT and missing data in at least one other baseline covariate. Precision of the odds ratios of the remaining baseline covariates is low — apparently because of the low number of missing data in general.

Patients in the PETAL trial intention-to-treat population have a median follow-up of 51.7 (95% CI: $[50.1; 53.4]$) months which corresponds to 74.2% of their potential follow-up time. Figure 5a shows the distribution of follow-up time and Figure 5b illustrates the individual follow-up time with respect to the time of the entry into the study.

Patient characteristics of the PETAL population are shown in Table 4 and the Kaplan-Meier curve for overall survival is depicted in Figure 6. Mean age at baseline is 58 years (median: 60 years) and 56% of the PETAL patients are males. The most frequent diagnosis is DLBCL and the majority of patients belong to the low risk group of the IPI_{Shipp}. The correlation between SUV at baseline and at interim PET/CT scan is 0.24 with Pearson's and 0.30 with Spearman's correlation coefficient. Within the documented follow-up time, 185 patients (21.5%) actually died. Median overall

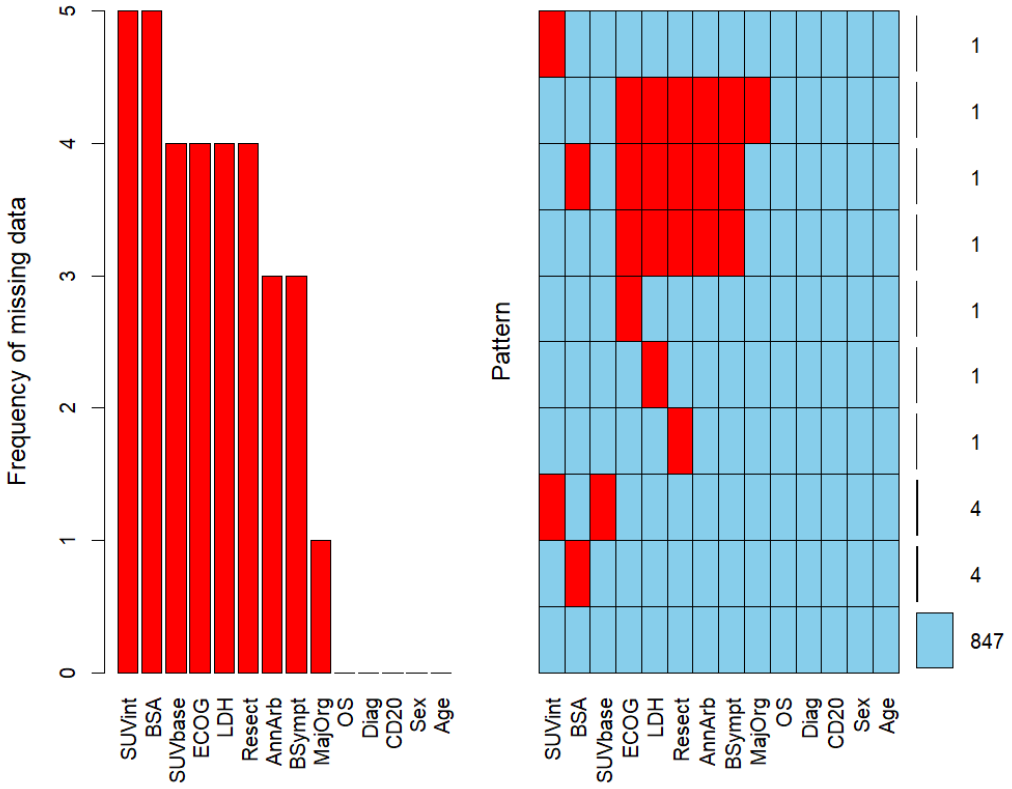
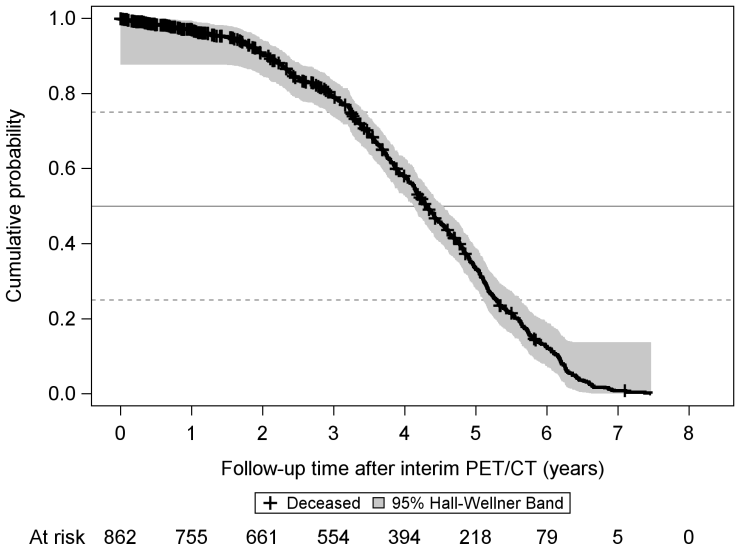
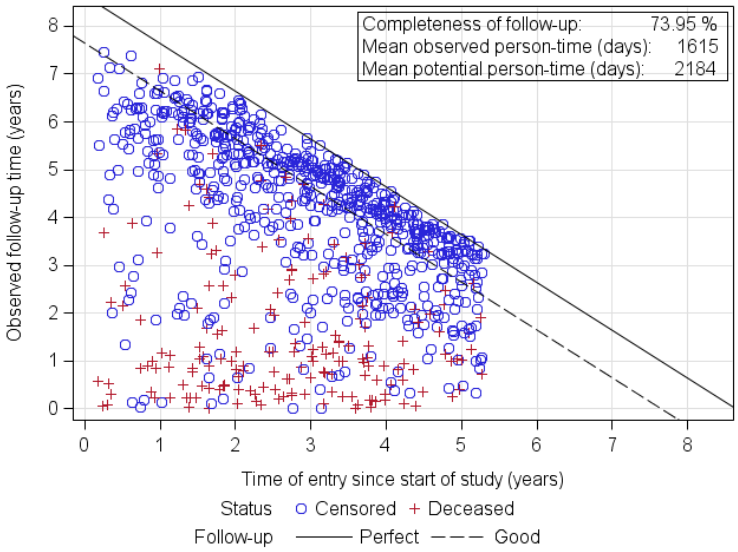


Figure 4: Frequency and pattern of missing data regarding the variables in the candidate set and overall survival (OS). Red colour indicates variables with missing data while blue colour represents variables with available data. Numbers on the right-hand side illustrate the number of patients with the respective pattern of available and missing data.



(a)



(b)

Figure 5: (a) Distribution of follow-up time estimated with the reverse Kaplan-Meier approach and (b) completeness of follow-up. Blue circles indicate patients who were alive at the end of their follow-up period and red crosses indicate deceased patients.

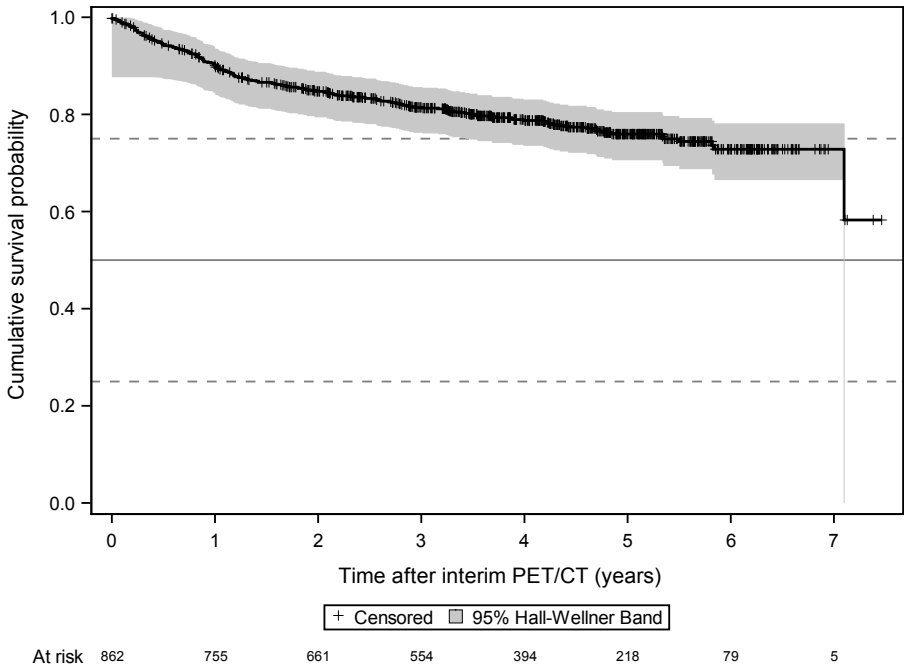


Figure 6: Kaplan-Meier plot for overall survival from interim PET/CT staging in the intention-to-treat population of the PETAL trial.

survival time from therapy allocation could, thus, not be determined as less than half of the patients died. The 25th percentile of overall survival time from therapy allocation, however, is 64.1 months and the 2-year overall survival Kaplan-Meier estimate is 0.847 (95% CI: [0.823; 0.872]). With respect to the time point of interest τ (two years after interim PET/CT), 73 patients were censored before τ while 128 died within the first two years and 661 were known to have survived for at least two years after interim PET/CT.

Likelihood ratio tests for competing models of ECOG and Ann Arbor staging reveal that ECOG can be used as a continuous variable in the candidate set and that Ann Arbor staging should better be addressed in categorical form (Table 9 in the Appendix). Using cumulative martingale residuals, maximum SUV at baseline PET/CT and body surface area show tendencies towards non-linearity (Figures 7a and 7b). Age and the continuous version of ECOG seem to violate the proportional hazards assumption of the Cox regression model as indicated by Schoenfeld residuals (Figures 7c and 7d). The remaining candidate variables appear unsuspicious regarding these assumptions (Figures 22, 23, and 24 in the Appendix).

Table 4: Clinical characteristics of the PETAL trial intention-to-treat population ($N = 862$). Represented as mean ± standard deviation (median [Q1; Q3]) for continuous variables and percentages for categorical variables.

Baseline covariate	Intention-to-treat population (N=862)
Age	57.52 ± 14.73 (60 [48; 70])
Sex	
Male	56.5 %
Female	43.5 %
Diag	
DLBCL	70.6 %
OthB	14.7 %
T	8.8 %
Other	5.8 %
SUVbase	20.84 ± 10.89 (19.65 [12.62; 26.80])
SUVint	4.42 ± 4.29 (3.40 [2.30; 4.80])
ECOG	
0	46.1 %
1	44.4 %
2	6.8 %
3	2.2 %
AnnArb	
I	17.5 %
II	22.6 %
III	23.2 %
IV	36.3 %
MajOrg	30.4 %
LDH	55.3 %
IPI_{Shipp}	
Low risk	38.2 %
Low-intermediate risk	26.1 %
High-intermediate risk	21.1 %
High risk	14.5 %
Missing	1
CD20	91.2 %
BSympt	30.6 %
Resect	11.1 %
BSA	1.91 ± 0.21 (1.92 [1.78; 2.01])

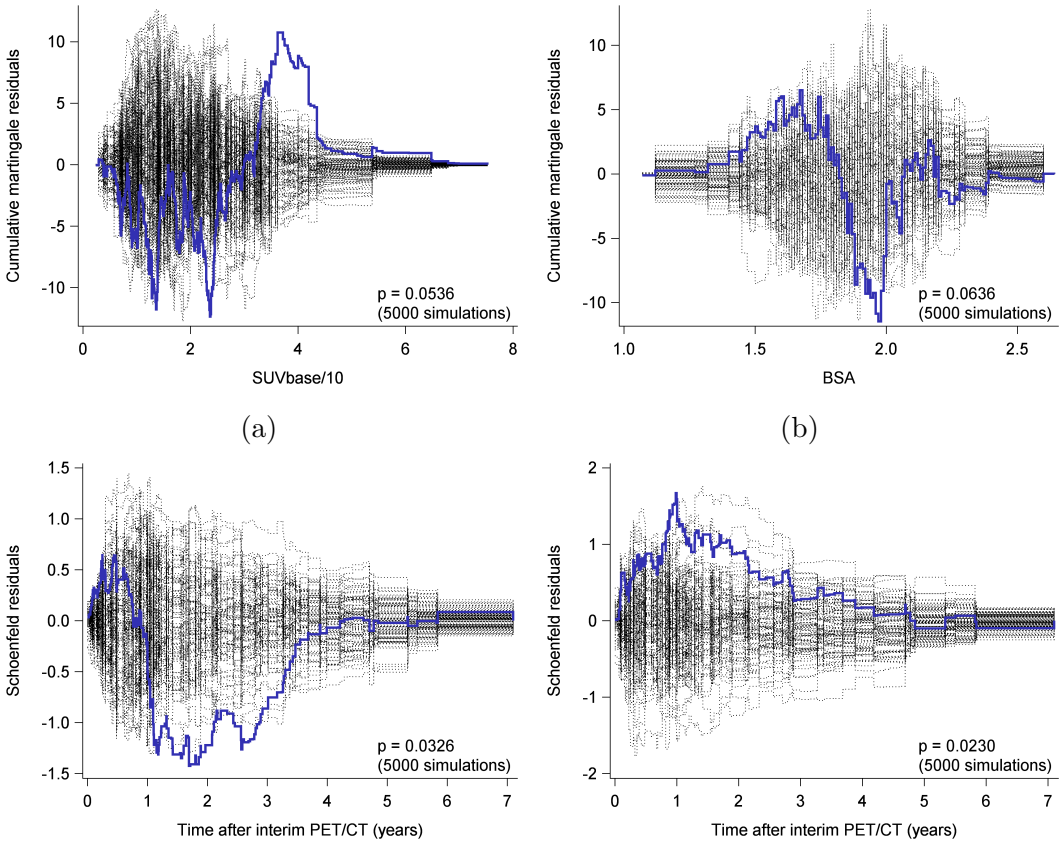


Figure 7: Checking the functional form for (a) maximum SUV at baseline PET/CT and (b) body surface area and checking the proportional hazards assumption for (c) age and (d) ECOG. The first fifty of $5,000$ simulated patterns are represented as dotted lines; the actually observed path is displayed as solid blue line.

3.2 Modelling-based prognostic risk score equations

This chapter presents the prognostic risk scores that are calculated from the results of the modelling approaches as described in Section 2.3. The three scores are somewhat similar and also the set of selected variables as well as the magnitude of their effects resemble each other. Backward variable selection leads to both maximum SUV at baseline and at interim PET/CT, age, LDH, and histological diagnosis being included in all three prognostic models. The logistic regression risk score also considers Ann Arbour stage while Cox regression selects ECOG and sex and the MFPT approach score additionally contains all three variables. Of note, the final model obtained by the MFPT approach includes non-linear fractional polynomial transformations of maximum SUV at baseline and at interim PET/CT but lacks any time-dependent effects. The resulting prognostic risk score models read as

$$\begin{aligned} R_{LogReg} = &- 0.423 \cdot (SUVbase/10) \, + \, 1.580 \cdot (SUVint/10) \\ &+ \, 0.490 \cdot (Age/10) \, + \, 0.720 \cdot \mathbf{I}_{(ULN,\infty)}(LDH) \\ &+ \, 0.004 \cdot \mathbf{I}_{\{II\}}(AnnArb) \, + \, 1.130 \cdot \mathbf{I}_{\{III\}}(AnnArb) \, + \, 1.099 \cdot \mathbf{I}_{\{IV\}}(AnnArb) \\ &- \, 0.981 \cdot \mathbf{I}_{\{OthB\}}(Diag) \, + \, 1.025 \cdot \mathbf{I}_{\{T\}}(Diag) \, - \, 0.627 \cdot \mathbf{I}_{\{Other\}}(Diag), \end{aligned}$$

$$\begin{aligned} R_{CoxReg} = &- 0.288 \cdot (SUVbase/10) \, + \, 0.715 \cdot (SUVint/10) \, + \, 0.441 \cdot (Age/10) \\ &+ \, 0.360 \cdot ECOG \, + \, 0.718 \cdot \mathbf{I}_{(ULN,\infty)}(LDH) \, - \, 0.496 \cdot \mathbf{I}_{\{female\}}(Sex) \\ &- \, 0.732 \cdot \mathbf{I}_{\{OthB\}}(Diag) \, + \, 0.689 \cdot \mathbf{I}_{\{T\}}(Diag) \, - \, 0.129 \cdot \mathbf{I}_{\{Other\}}(Diag), \end{aligned}$$

and

$$\begin{aligned} R_{MFPT} = &- 0.018 \cdot (SUVbase/10)^3 \, + \, 1.963 \cdot (SUVint/10)^{0.5} \, + \, 0.422 \cdot (Age/10) \\ &+ \, 0.276 \cdot ECOG \, + \, 0.539 \cdot \mathbf{I}_{(ULN,\infty)}(LDH) \, - \, 0.417 \cdot \mathbf{I}_{\{female\}}(Sex) \\ &+ \, 0.018 \cdot \mathbf{I}_{\{II\}}(AnnArb) \, + \, 0.522 \cdot \mathbf{I}_{\{III\}}(AnnArb) \, + \, 0.423 \cdot \mathbf{I}_{\{IV\}}(AnnArb) \\ &- \, 0.711 \cdot \mathbf{I}_{\{OthB\}}(Diag) \, + \, 0.753 \cdot \mathbf{I}_{\{T\}}(Diag) \, - \, 0.070 \cdot \mathbf{I}_{\{Other\}}(Diag). \end{aligned}$$

Note that age, maximum SUV at baseline PET/CT, and maximum SUV at interim PET/CT are divided by ten to have more interpretable effect sizes in the risk score formulations when rounding to three decimal places. Figure 8 gives a graphical overview with regard to the similarity of the modelling-based risk scores while Pearson's correlation coefficient is 0.867 (95% CI: $[0.850; 0.883]$) between the logistic regression and Cox regression risk scores, 0.887 (95% CI: $[0.872; 0.901]$) between logistic regression and MFPT approach, and 0.906 (95% CI: $[0.893; 0.917]$) between Cox regression and MFPT approach.

3.3 Discrimination and calibration performance

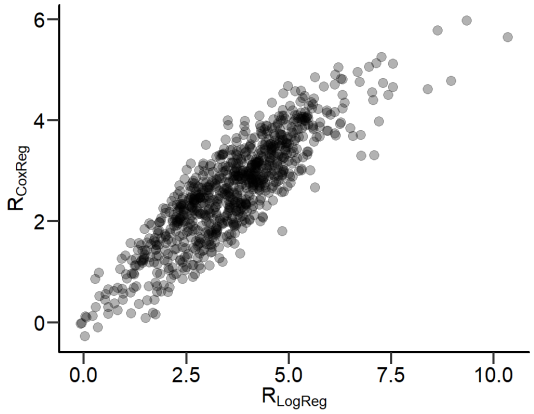
The optimal cut-off values of the prognostic risk scores according to the local discrimination measures defined in Section 2.4 are used to split the data into a good and a bad prognosis group. Those cut-offs are shown in Table 5 by local discrimination measure and prognostic risk score. For the IPI-based scores as well as for the logistic regression and MFPT approach risk scores, the criteria closest-to-(0,1) corner, Youden index, and concordance probability lead to one and the same cut-off point. Cox regression realises a slightly smaller cut-off point according to Euclidean distance than according to the other two local discrimination measures.

Table 5: Optimal cut-off points by local discrimination measure (ED: closest-to-(0,1) corner criterion; J: Youden index; CP: concordance probability) and prognostic risk score (LogReg: logistic regression; CoxReg: Cox regression; MFPT: MFPT approach).

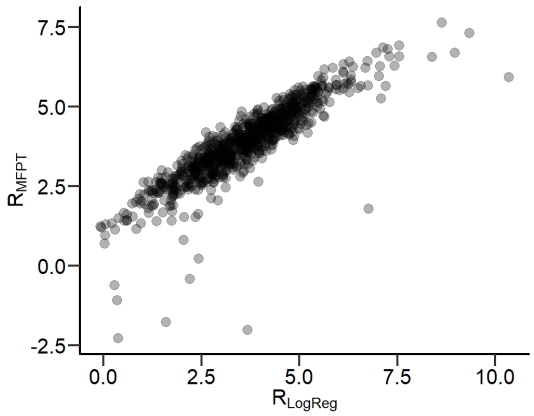
	Optimal cut-off* point according to		
	ED	J	CP
IPI_{Shipp}	2	2	2
IPI_{iPET}	2	2	2
LogReg	4.29	4.29	4.29
CoxReg	2.86	3.28	3.28
MFPT	4.46	4.46	4.46

*Exceeding the respective optimal cut-off point defines the bad prognosis group.

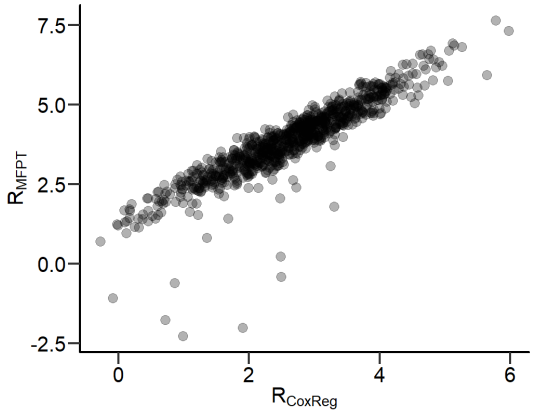
The ROC curves depicted in Figure 9a give an overall impression of the discrimination performance of the five risk scores. Corresponding AUC values and results according to the local discrimination measures defined in Section 2.4 are summarised in Figure 10 while detailed numerical results are available in Table 12 in the Appendix. Remember



(a)

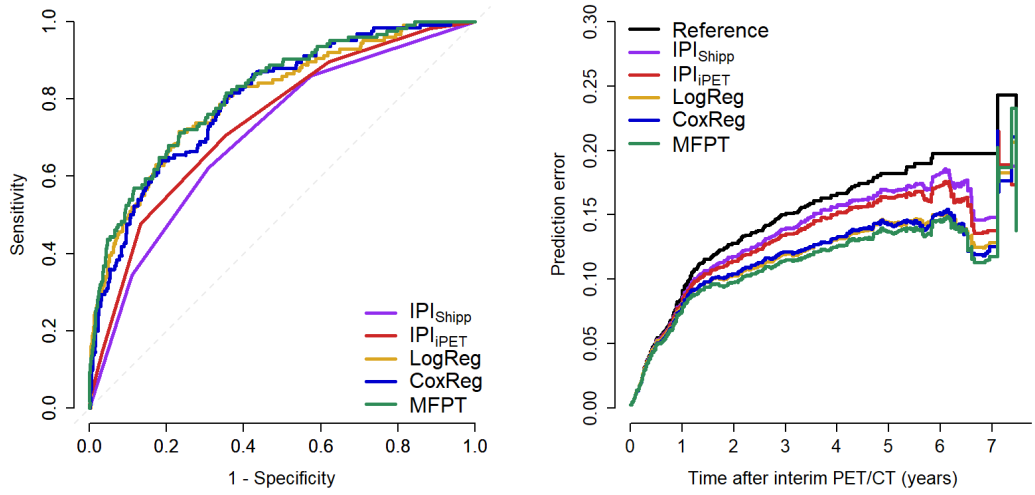


(b)



(c)

Figure 8: Scatter plots of (a) the logistic (R_{LogReg}) and Cox regression (R_{CoxReg}) risk scores, (b) the logistic regression and MFPT approach (R_{MFPT}) risk scores, (c) and the Cox regression and MFPT approach risk scores.



(a) (b)

Figure 9: (a) Receiver operating characteristic curves and (b) prediction error curves for overall survival by prognostic risk score. A prediction error curve is also plotted for the marginal Kaplan-Meier prediction model that is indicated as reference model.

here that for all measures but for the Euclidean distance higher values indicate better discrimination. One can see that irrespective of the measure used the three scores obtained by the modelling approaches logistic regression, Cox regression, and MFPT approach show better performance than the integer-based scores which is also in accordance with the ROC curves. Examining the results in more detail — for example with regard to the AUC — logistic regression ($AUC_{LogReg} = 0.802$, 95% CI: $[0.755; 0.850]$) and its modelling competitors Cox regression ($AUC_{CoxReg} = 0.800$, 95% CI: $[0.752; 0.848]$) and MFPT approach ($AUC_{MFPT} = 0.817$, 95% CI: $[0.771; 0.863]$) are almost on one and the same level with a slightly better performance of the MFPT approach. Within the IPI-based scores, IPI_{iPET} ($AUC_{IPI_{iPET}} = 0.741$, 95% CI: $[0.690; 0.793]$) seems to outperform the original IPI_{Shipp} ($AUC_{IPI_{Shipp}} = 0.708$, 95% CI: $[0.655; 0.760]$) to a small extent.

Model calibration as measured by the Brier score over time seems to differ between the IPI-based scores and the three scores obtained by regression models. As can be seen from Figure 9b, the IPI-based prediction error curves lie between the curve of the reference model and those of logistic regression, Cox regression, and the MFPT approach. Here, the reference model refers to the marginal Kaplan-Meier prediction model ignoring any predictor variables. The integrated Brier score as summarising measure accordingly confirms this tendency with the best results for MFPT ($IBS_{MFPT} = 0.113$) followed by

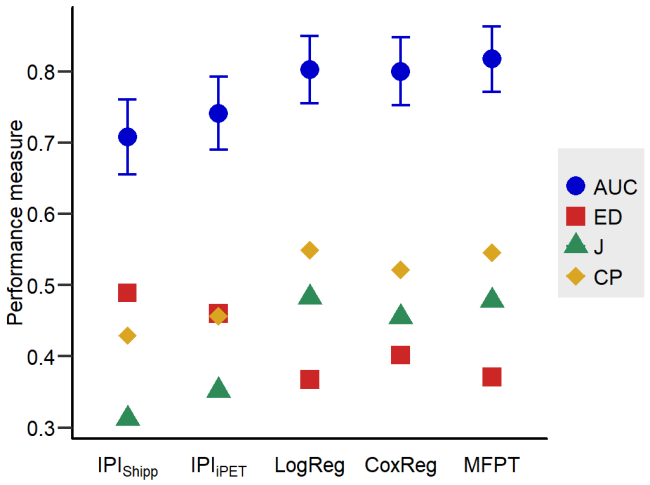


Figure 10: Area under the curve (AUC) with 95% confidence interval and local discrimination measures closest-to-(0,1) corner criterion (ED), Youden index (J), and concordance probability (CP) by prognostic risk score (LogReg: logistic regression; CoxReg: Cox regression; MFPT: MFPT approach).

Cox regression ($IBS_{CoxReg} = 0.117$) and logistic regression ($IBS_{LogReg} = 0.118$). While IPI_{Shipp} ($IBS_{IPI_{Shipp}} = 0.136$) and IPI_{iPET} ($IBS_{IPI_{iPET}} = 0.131$) still outperform the reference model ($IBS_{Reference} = 0.151$), their results are considerably worse than those of the modelling approaches.

When using the scores' optimal cut-off points with regard to the closest-to-(0,1) corner criterion to differentiate between patients with bad and good prognosis under (R-)CHOP therapy, logistic regression and the MFPT approach classify the least patients into the bad prognosis group (261/853). IPI_{Shipp} (307/861) allocates more patients to that group whereas IPI_{iPET} (347/853) and Cox regression (351/852) are on a even higher level. To give an impression of the similarity of the scores regarding the classification into bad and good prognosis, Table 6 provides two-way tabulations on a score-by-score basis. Additionally, Table 10 (in the Appendix) shows the resulting concordance between the scores' classifications, and Table 11 (also in the Appendix) relates the five risk scores' prognostic groups to the survival status at time point τ.

In Figure 11, Kaplan-Meier curves by prognosis according to the respective optimal cut-off point with respect to the closest-to-(0,1) corner criterion are shown for IPI_{Shipp}, IPI_{iPET}, logistic regression, Cox regression, and MFPT approach. Graphically, the best segregation can be found for Cox regression which is associated with a hazard ratio of 5.807 (95% CI: $[4.122; 8.181]$). On the other side of the medal, the integer-based

Table 6: Cross-tables of the five risk scores regarding the classification into good and bad prognosis patient groups based on the closest-to-(0,1) corner criterion.

		IPI_{iPET}		LogReg		CoxReg		MFPT	
		Good	Bad	Good	Bad	Good	Bad	Good	Bad
IPI_{Shipp}	Good	506	44	462	88	395	154	451	98
	Bad	0	303	130	173	106	197	140	163
IPI_{iPET}	Good			448	58	382	123	438	67
	Bad			144	203	119	228	153	194
LogReg	Good					469	122	544	47
	Bad					32	229	47	214
CoxReg	Good							491	10
	Bad							100	251

scores IPI_{Shipp} and IPI_{iPET} perform worst here with hazard ratios of 2.593 (95% CI: $[1.938; 3.469]$) and 2.980 (95% CI: $[2.201; 4.035]$), respectively. Nevertheless, they still seem to provide reasonable separation of the survival curves.

3.4 Internal validation

8-fold cross-validation This section presents the results of the 8-fold internal cross validation that is used to estimate the robustness of the results and to identify possible overoptimism of the modelling approaches. Figure 12 illustrates the separation of the PETAL data into eight validation sample sub-data sets of equal size with respect to the date of treatment allocation after interim PET/CT. The first treatment allocation took place on the 29th of January 2008 and the last in 2013 on the 15th of February. Recruitment started slowly with the treatment allocation dates of the first validation sample covering about one year but accelerating and stabilising thereafter. While IPI_{Shipp} and IPI_{iPET} have their pre-defined set of variables, selection of variables from the candidate set varies by modelling approach. Table 7 shows the frequency of the modelling approaches selecting the variables from the candidate set into the final model. It can be seen that age and SUV at interim PET/CT are included in each of the eight validation samples for all three approaches. Of note, SUV at interim PET/CT is selected more often than SUV at baseline PET/CT. Lymphomatous involvement in major organs, body surface area, and resected manifestations are not included in any final model while selection frequencies for ECOG, Ann Arbor staging, LDH, and sex vary by the prognostic model used. Histological diagnosis appears in all but two final models and CD20 expression as well as B symptoms in only one.

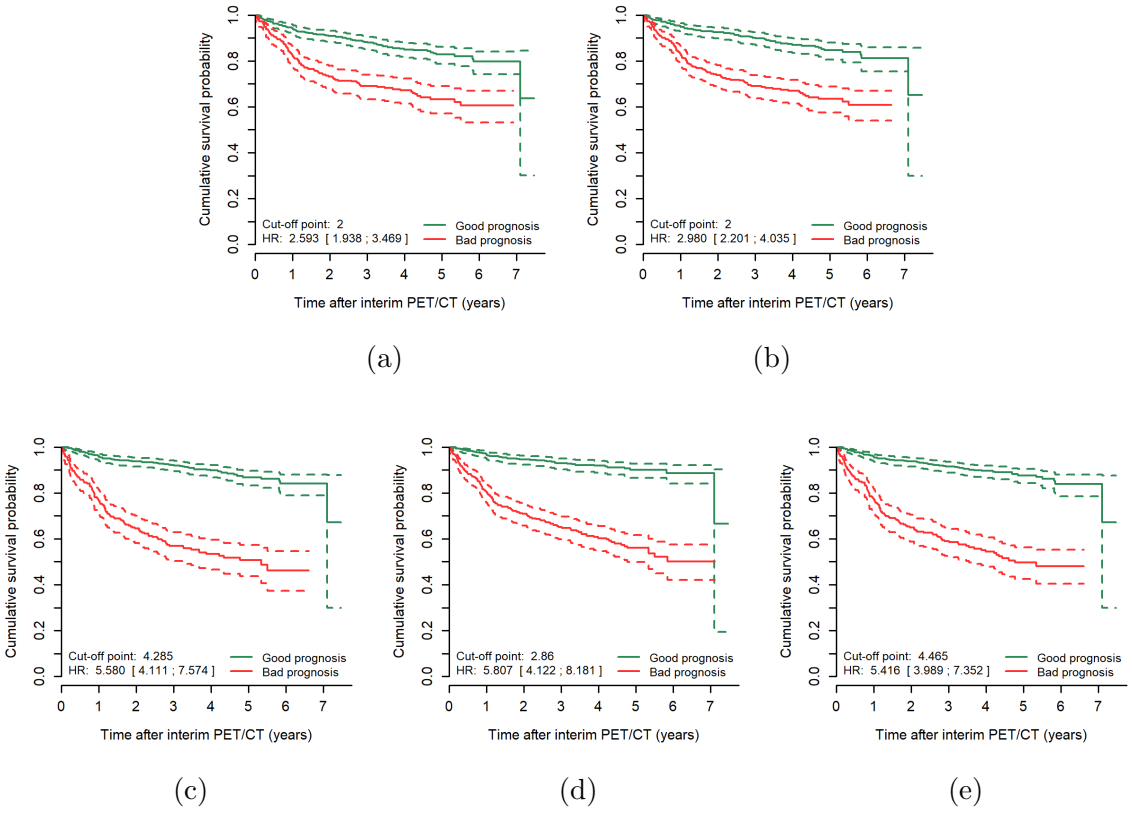


Figure 11: Kaplan-Meier curves for overall survival by prognosis according to the cut-off points obtained by the closest-to-(0,1) corner criterion for (a) IPI_{Shipp}, (b) IPI_{iPET}, (c) logistic regression, (d) Cox regression, and (e) MFPT approach.

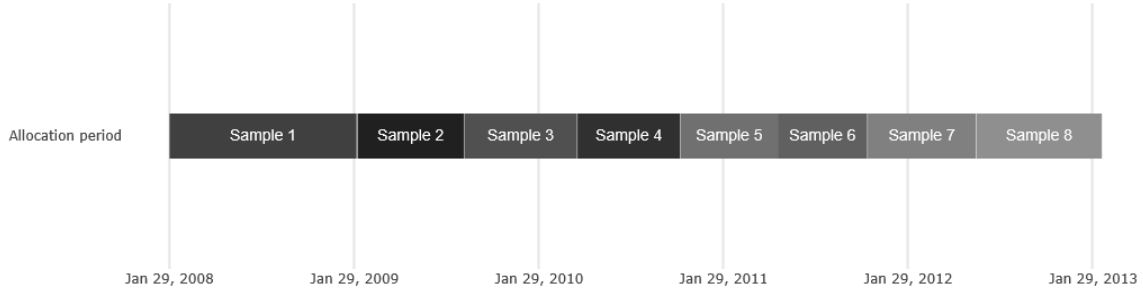


Figure 12: Treatment allocation dates by validation sample for the PETAL data. Vertical lines indicate full years after the first treatment allocation during the study.

Table 7: Absolute variable selection frequency by modelling approach in the 8-fold cross validation. Maximum number of selections of a variable for logistic regression (LogReg), Cox regression (CoxReg), and MFPT approach (MFPT) is eight.

Variable	LogReg	CoxReg	MFPT
Age	8	8	8
ECOG	2	7	4
AnnArb	7	1	4
LDH	4	8	8
MajOrg	0	0	0
Sex	3	6	4
Diag	7	8	7
CD20	0	0	1
BSympt	1	0	0
BSA	0	0	0
Resect	0	0	0
SUVbase	6	7	6
SUVint	8	8	8

The results by validation sample for area under the curve, Euclidean distance, Youden index, and concordance probability are shown in Figure 13. The advantage in discrimination performance of the modelling approaches over the integer-based scores appears to be lower than in the entire data set. Still, the tendency of IPI_{iPET} outperforming IPI_{Shipp} seems to be confirmed. Across all measures, the MFPT approach produces the most compact results without any notable outliers. Evidently, one specific validation sample poses challenges to all five scores.

ROC curves, prediction error curves, and hazard ratios by prognostic risk score and validation sample are shown in Figures 14, 15, and 16, respectively. At least the ROC curves and the hazard ratio figures are in line with the findings for the AUC and the local discrimination measures that the validation results for MFPT seem to be the most robust among the five prognostic risk scores.

Validation sample with miserable performance in the test data set This subsection deals with validation sample number four that experiences bad prognostic performance for all five risk scores. Associated ROC curves and prediction error curves are shown in Figure 17. Corresponding AUC values reach from 0.536 (95% CI: $[0.380; 0.693]$) for IPI_{Shipp} to 0.621 (95% CI: $[0.464; 0.778]$) for the MFPT approach. IBS values scatter around 0.140 ($IBS_{Reference} = 0.144$) for all risk scores with MFPT

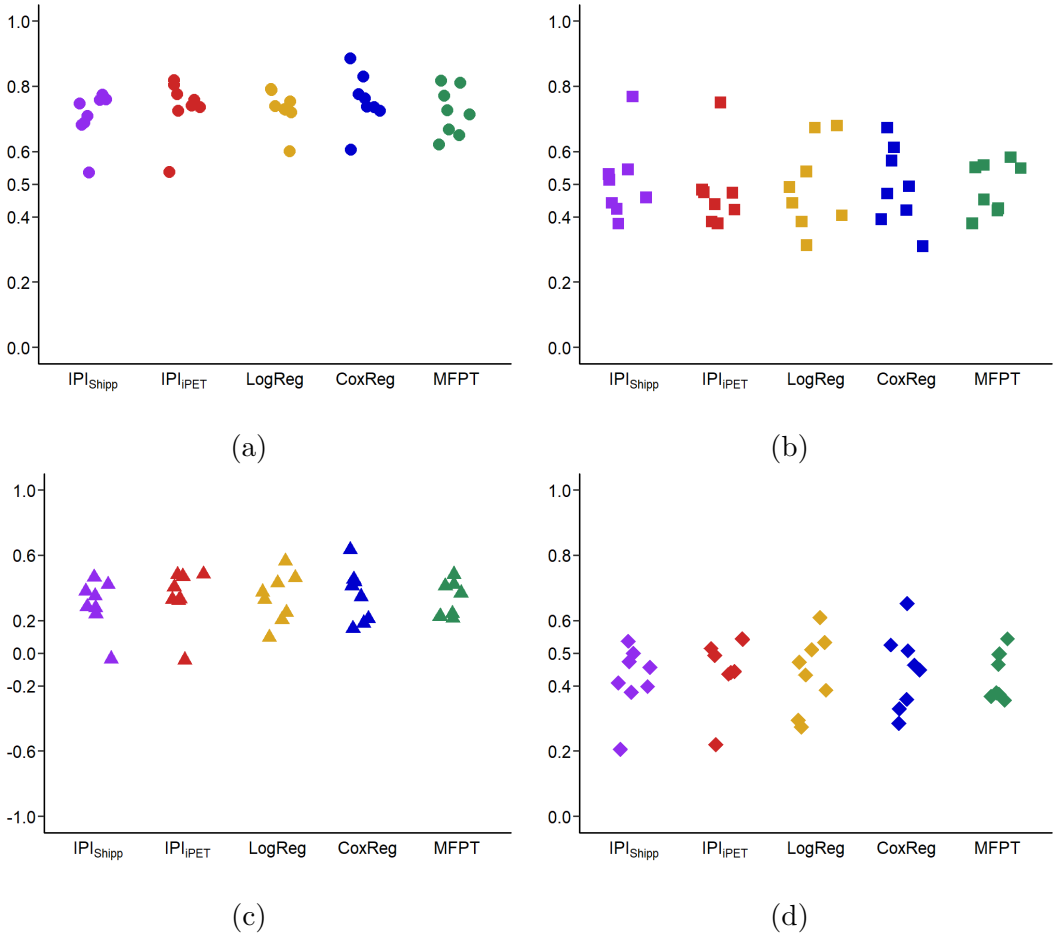


Figure 13: Results of the 8-fold cross validation for (a) area under the curve, (b) Euclidean distance, (c) Youden index, and (d) concordance probability by prognostic risk score (LogReg: logistic regression; CoxReg: Cox regression; MFPT: MFPT approach).

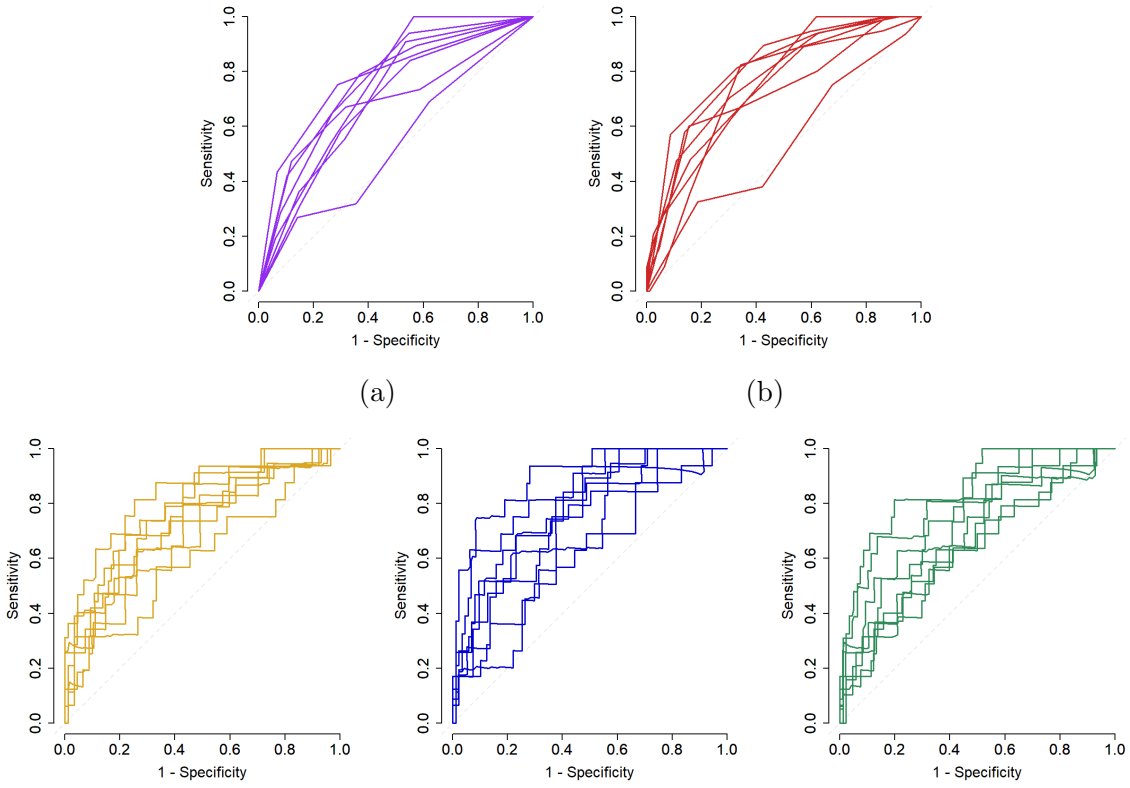


Figure 14: Receiver operating curves by validation sample for (a) IPI_{Shipp}, (b) IPI_{iPET}, (c) logistic regression, (d) Cox regression, and (e) MFPT approach.

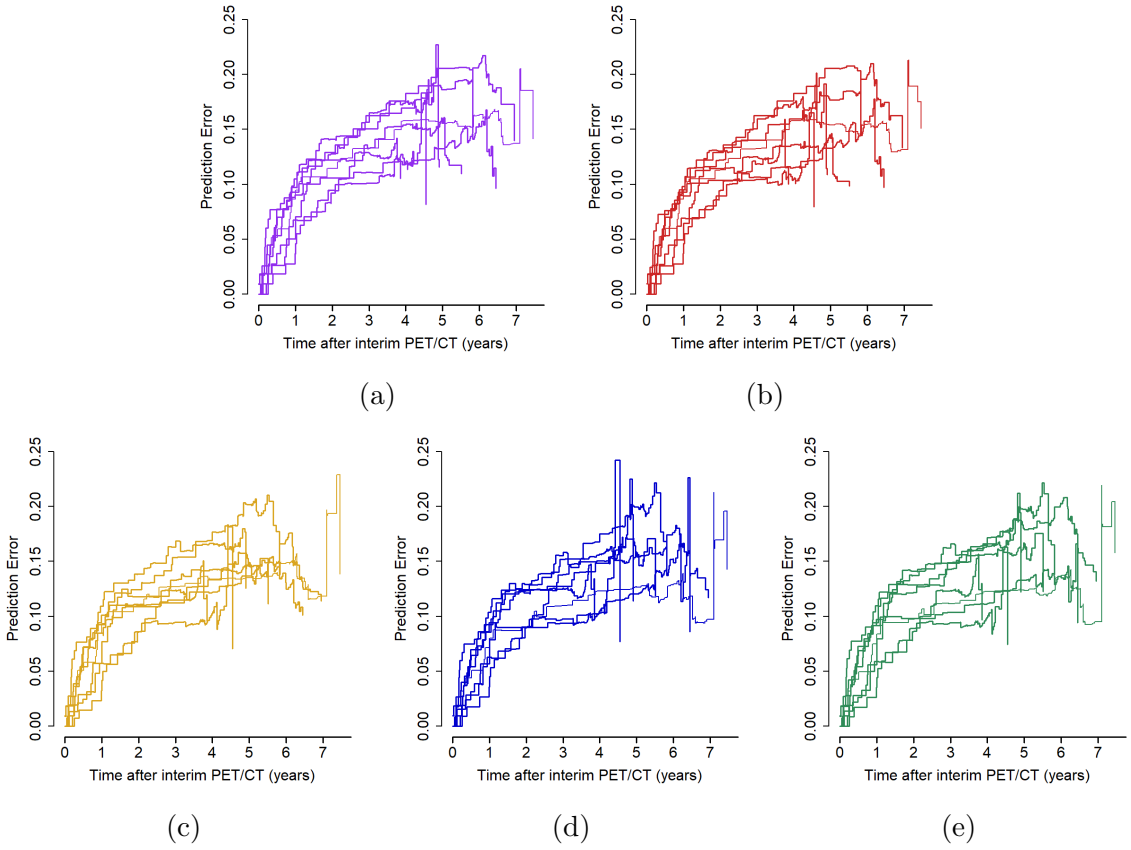


Figure 15: Prediction error curves by validation sample for (a) IPI_{Shipp}, (b) IPI_{iPET}, (c) logistic regression, (d) Cox regression, and (e) MFPT approach.

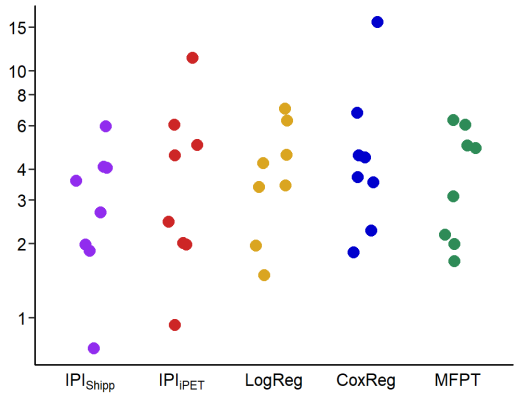
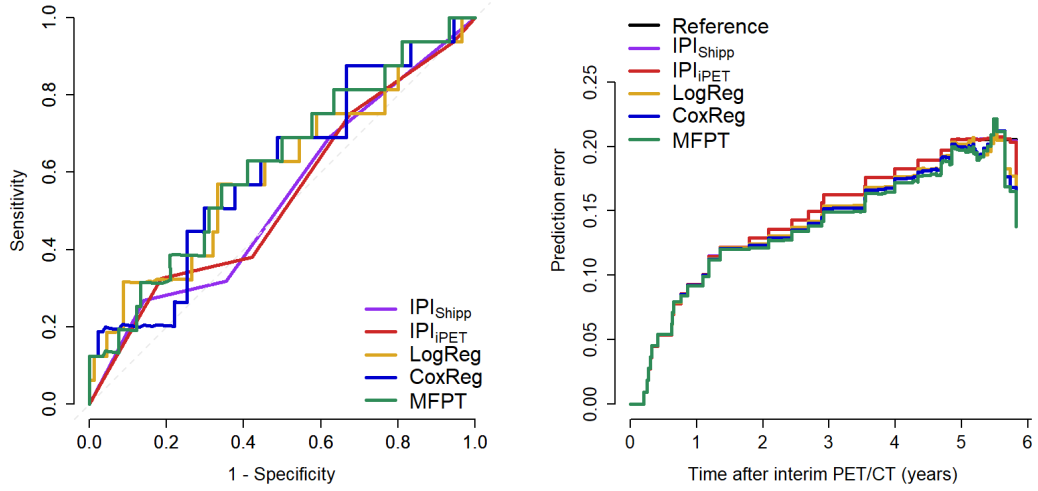


Figure 16: Hazard ratios for overall survival between bad and good prognosis patients (according to the closest-to-(0,1) corner criterion) by validation sample for IPI_{Shipp}, IPI_{iPET}, logistic regression (LogReg), Cox regression (Cox regression), and MFPT approach (MFPT).



(a) (b)

Figure 17: (a) Receiver operating characteristic curves and (b) prediction error curves by prognostic risk score (LogReg: logistic regression; CoxReg: Cox regression; MFPT: MFPT approach) in the test data set of validation sample four.

performing best with regard to this metric ($IBS_{MFPT} = 0.136$). Hazard ratios for overall survival between patients with bad prognosis and patients with good prognosis as indicated by the optimal cut-off point according to the closest-to-(0,1) corner criterion are also low (Figure 18). Of note, for IPI_{Shipp} and IPI_{iPET}, bad prognosis patients even show better survival than good prognosis patients. It seems worthwhile to take a closer look at how the candidate variables affect the risk of dying in the training and test data, respectively. Regarding the clinical characteristics of the patients, there do not seem to be remarkable differences between training and test data set (Table 13 in the Appendix). However, in a Cox regression model for overall survival with SUV at baseline and at interim PET/CT, age, LDH, Ann Arbour stage, and histological diagnosis as prognostic variables, the hazard ratios of some variables differ between the training and the test data. That is, in contrast to the training data set, a higher LDH seems to be protective in the test data set and also stage I seems to be associated with the highest risk of dying among the four Ann Arbor stages (Table 14 in the Appendix).

Validation sample selecting a time-dependent effect in the training data set
Another remarkable validation sample is number two as it is the only one to have a time-dependent effect selected in the training data set by the MFPT approach. The

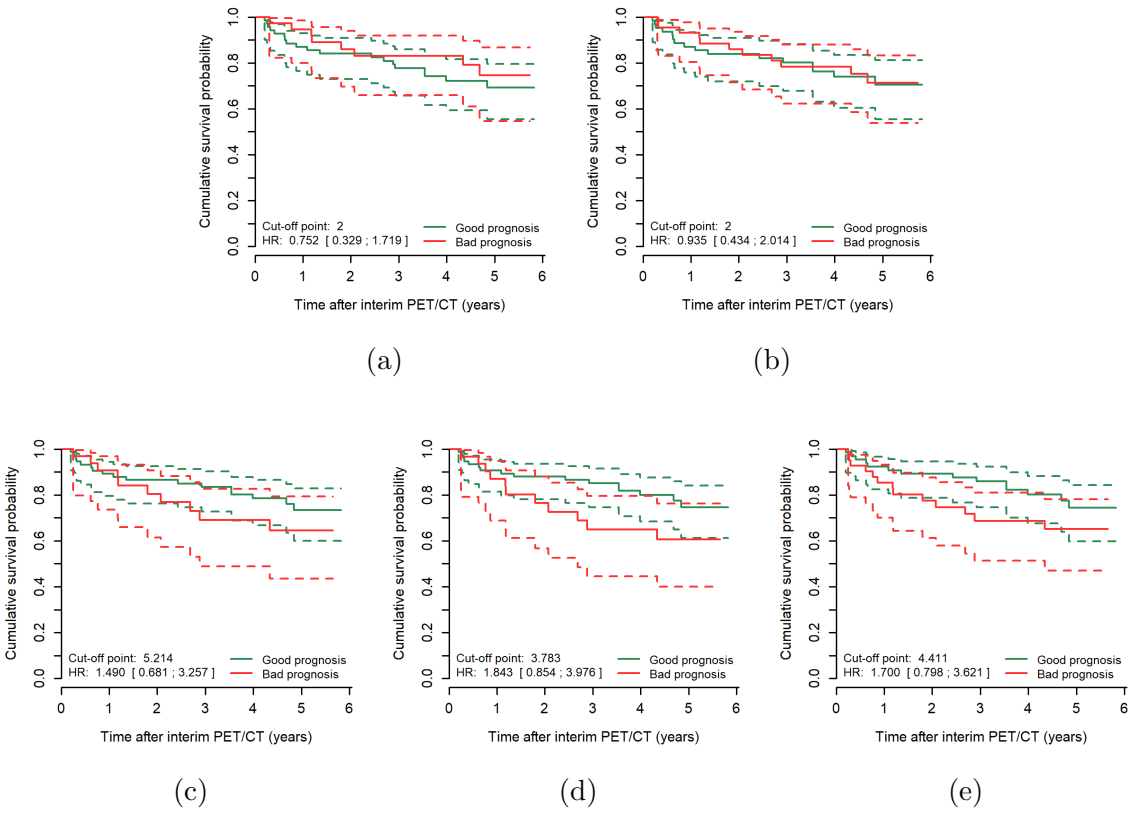


Figure 18: Kaplan-Meier curves for overall survival by prognosis according to the cut-off points obtained by the closest-to-(0,1) corner criterion for (a) IPI_{Shipp}, (b) IPI_{iPET}, (c) logistic regression, (d) Cox regression, and (e) MFPT approach in the test data set of validation sample four.

MFPT score for this validation sample reads as

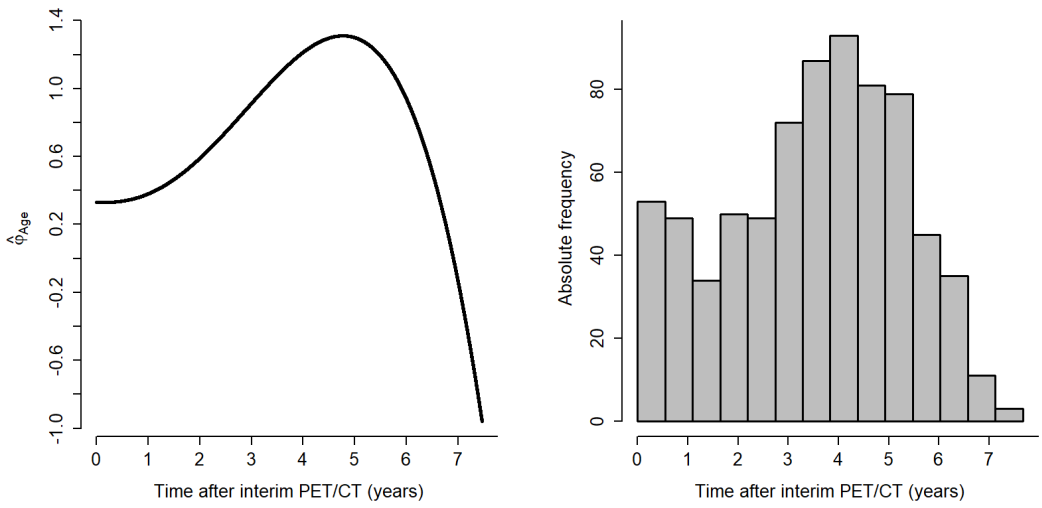
$$\begin{aligned} R^{Val2}_{MFPT} = & -0.213 \cdot (SUVbase/10) \\ & + 0.031 \cdot ((SUVint/10)^2 + (SUVint/10)^2 \cdot \log((SUVint/10)) \\ & + 0.962 \cdot (Age/10) + 0.679 \cdot \mathbf{I}_{(ULN,\infty)}(LDH) \\ & - 0.689 \cdot \mathbf{I}_{\{OthB\}}(Diag) + 0.885 \cdot \mathbf{I}_{\{T\}}(Diag) - 0.562 \cdot \mathbf{I}_{\{Other\}}(Diag) \\ & - 0.036 \cdot \mathbf{I}_{\{II\}}(AnnArb) + 0.625 \cdot \mathbf{I}_{\{III\}}(AnnArb) + 0.604 \cdot \mathbf{I}_{\{IV\}}(AnnArb). \end{aligned}$$

The final weight ω_{Age} for age is elicited from the time-dependent function

$$\widehat{\varphi}_{Age}(t, (q_1 = 3, q_2 = 3)) = 0.3293043 + 4.35 \cdot 10^{-9} \cdot t^3 - 5.58 \cdot 10^{-10} \cdot t^3 \cdot \log(t)$$

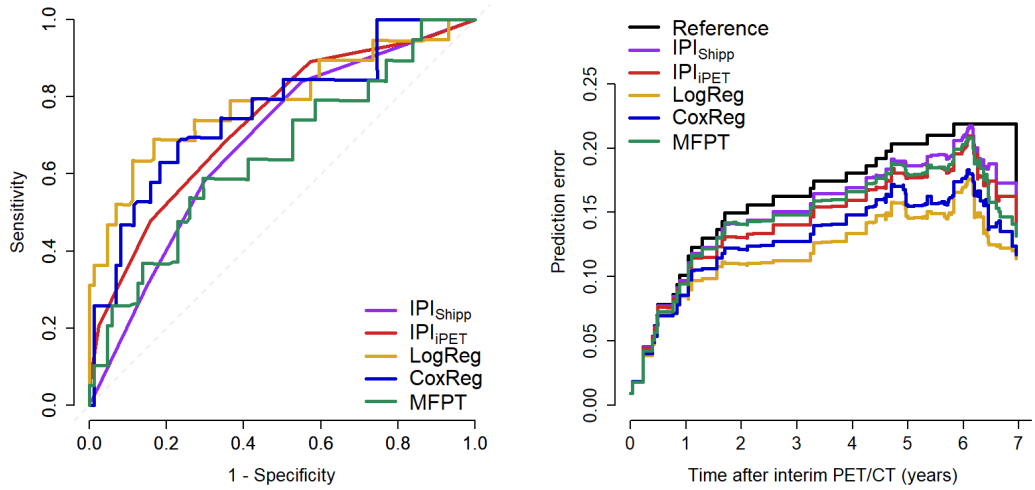
as described in Section 2.3 where t is the time component and given as days after interim PET/CT. The weight for age in its time-dependent form and the distribution of event times in the training sample that are used to summarise the time-dependent function into one single number are displayed in Figure 19. The variables included in the MFPT model are the same as in the entire PETAL data set apart from the fact that ECOG and sex are missing. Also, the magnitudes of the effects are comparable with only the weight for age derived from the time-dependent function deviating considerably — 0.962 here and 0.422 in the entire data set.

Prognostic performance of the MFPT-based risk score, however, is poor compared to the initial analysis results in the entire PETAL data set. Associated ROC curves and prediction error curves are shown in Figure 20. The AUC for the MFPT approach is 0.651 (95% CI: $[0.506; 0.795]$) while IPI_{Shipp} ($AUC_{IPIShipp} = 0.683$, 95% CI: $[0.540; 0.825]$), IPI_{iPET} ($AUC_{IPIiPET} = 0.736$, 95% CI: $[0.600; 0.873]$), logistic regression ($AUC_{LogReg} = 0.789$, 95% CI: $[0.661; 0.916]$), and Cox regression ($AUC_{CoxReg} = 0.763$, 95% CI: $[0.630; 0.895]$) show considerably better performance. The IBS is better for the MFPT approach than for the IPI_{Shipp} but worse than for the three remaining risk scores. For the hazard ratio between bad and good prognosis patients, MFPT is clearly outperformed by logistic regression and Cox regression and is on a similar level as the IPI-based scores (Figure 21).



(a) (b)

Figure 19: (a) Time-dependent weighting function for age and (b) distribution of event times both elicited from the training data set of validation sample two.



(a) (b)

Figure 20: (a) Receiver operating characteristic curves and (b) prediction error curves by prognostic risk score (LogReg: logistic regression; CoxReg: Cox regression; MFPT: MFPT approach) in the test data set of validation sample two.

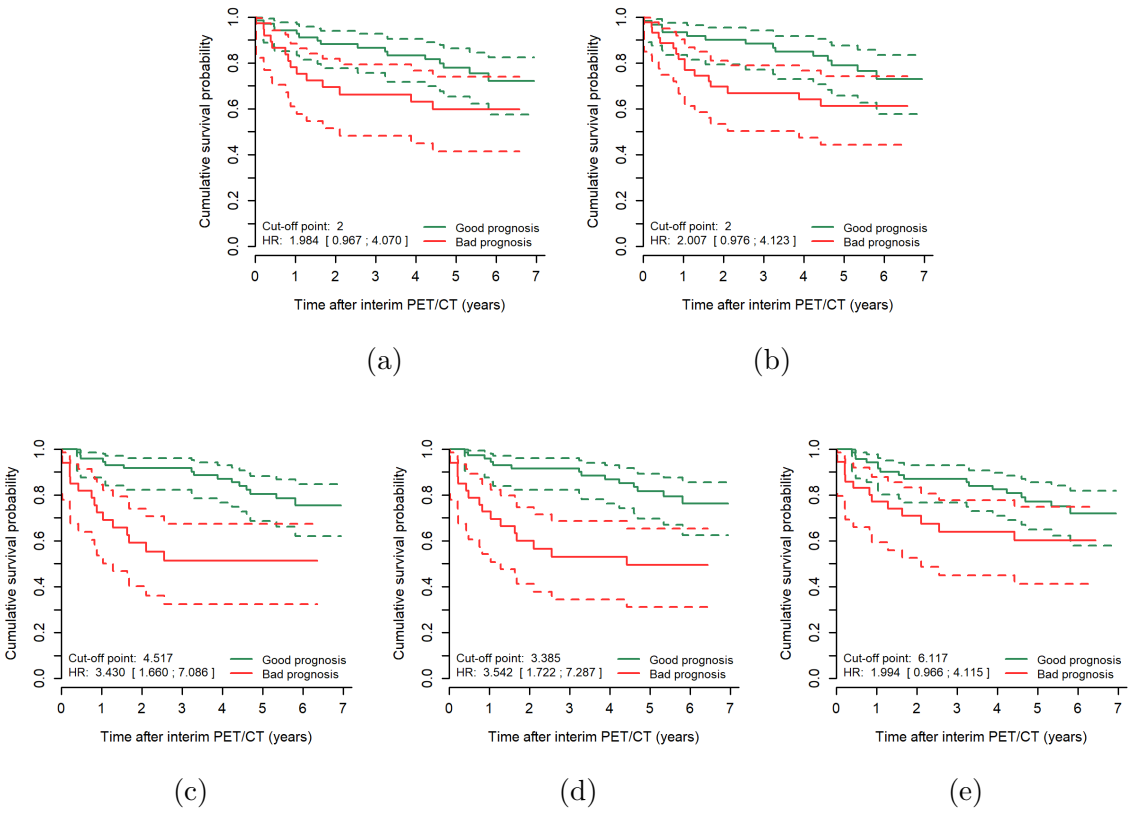


Figure 21: Kaplan-Meier curves for overall survival by prognosis according to the cut-off points obtained by the closest-to-(0,1) corner criterion for (a) IPI_{Shipp}, (b) IPI_{iPET}, (c) logistic regression, (d) Cox regression, and (e) MFPT approach in the test data set of validation sample two.

4 Discussion

The need for a prognostic risk model The general need to develop a prognostic risk model for aggressive NHL under (R-)CHOP raises from the major drawback of exclusively using the 66% cut-off point for the deltaSUVmax method (Lin et al., 2007) as, for example, employed in the PETAL trial (Dührsen et al., 2018). That is, the low fraction ($108/862 = 12.5\%$) of patients with bad prognosis according to this criterion led to failing the recruitment aim in the interim PET/CT positive part of the study. The results of this thesis indicate that the issue can be solved by building risk scores from multivariable prognostic models as the proportions of bad prognosis patients for the scores obtained from logistic regression, Cox regression, and MFPT approach are considerably higher than for the 66% cut-off point alone. Additionally, they even offer better segregation of survival curves with respect to hazard ratios (cf. Dührsen et al., 2018, Supplemental Figure A2, Panel B). The crucial question whether a modelling-based risk score makes it into clinical practice depends, however, to a great extent on its practicality. In general, prognostic "models presented in the literature are a compromise between the extremes of the 'statistical ideal' and the 'clinical ideal'" (van Houwelingen, 2000). That is, an optimal prognostic model from statistical point of view may provide great goodness of fit but from clinical point of view may include too many variables with some of them being difficult or expensive to obtain. In contrast, the IPI_{Shipp} may serve as a role model for the clinical ideal. From statistical perspective, the major drawback of the IPI_{Shipp} is its simple construction with one penalty point for each risk factor (van Houwelingen, 2000). Although the dichotomisation of its factors comes at the price of loss of information (Royston et al., 2006), this simplicity may be one of the reasons why the IPI_{Shipp} has retained its value for clinicians all over the world for almost thirty years as it makes the IPI_{Shipp} easy to use in everyday practice. In times of the computer era, however, simplicity of a prognostic risk score is not necessarily needed as each model can easily be translated into and displayed as a nomogram, a spreadsheet, a web-based computer programme, or a smartphone application (Moons et al., 2012). Nevertheless, a successful prognostic risk score still must be accepted and used by treating physicians and needs to be supported by the clinical community.

General conclusions The main question of this thesis is whether it is possible to develop a prognostic risk score for aggressive NHL patients under (R-)CHOP therapy that includes early response to therapy and outperforms the IPI_{Shipp} in terms of

discrimination and calibration. The results of this thesis show that the interim PET/CT-enhanced IPI is superior to the IPI_{Shipp} with respect to the AUC, the Euclidean distance, and the Brier score, and thus, such score can be constructed. Consequently, early response to (R-)CHOP therapy measured as ΔSUV has additional prognostic value beyond the IPI_{Shipp} in aggressive NHL patients. Whether scores obtained from more elaborated prognostic models are superior to the integer-based IPI scores, remains unclear but at least there is a tendency towards better performance of the model-build scores. Most likely, the advantage of the modelling-based over the integer-based scores in the entire data set is due to the loss of information associated with the dichotomisation of continuous variables (Royston et al., 2006) that is inherent in the IPI_{Shipp} and the IPI_{iPET}. Additionally, it may be due to more prognostic factors being taking into account by the logistic regression, Cox regression, and MFPT approach scores, e.g., the histological diagnosis. However, one should also safeguard against the potential of overfitting that may come along with the modelling approaches and may lead to an overoptimism in the corresponding discrimination measure estimates (Steyerberg, 2009). Using cross-validation, superiority of the modelling approaches over the integer-based scores in terms of discrimination performance appears to be lower than initially. Thus, overoptimism due to the model-building process seems at least possible. Before any of the proposed scores developed with the PETAL data can be implemented in clinical practice, the extent of their overoptimism needs to be estimated reliably. Therefore, an external validation study is needed to assess their discrimination and calibration performance in an independent patient population in order to prove their generalisability. With respect to the third level of investigation as to whether non-linearity and non-proportional hazards should be considered in the development of an aggressive NHL risk score, the MFPT approach seems to provide more robust results in internal cross-validation than logistic and Cox regression. This robustness is also reflected by the fact that the MFPT approach (as well as logistic regression) leads to one and the same cut-off value irrespective of the choice of the local discrimination measure. The final MFPT model, however, only considers non-linearity but does not reject proportionality of the hazards for any variable. The former is interestingly the case for both SUV variables where only for baseline PET/CT possible deviances from linearity are indicated by cumulative martingale residuals. While tendencies towards non-proportional hazards at least for age and ECOG motivate the use of the MFPT approach, it neither selects a time-dependent effect for these two nor for any other variables in the final model. One explanation may be that the proportional hazards assessment is adjusted for all candidate variables

whereas step three of the MFPT approach only addresses the variables in model $\mathcal{M}_1$ selected in the first and second steps of the procedure. Here, a simulation study seems worthwhile to elicit realistic aggressive NHL scenarios that benefit the most from the time-dependent approach compared to standard Cox regression. Such a study could include the investigation of the extent of violations of linearity and proportional hazards assumptions either separately or in combination.

A closer look from clinical perspective As to the variables selected by the prognostic modelling approaches, the three scores share SUV at baseline and at interim PET/CT, age, LDH, and histological diagnosis. While the logistic regression score contains Ann Arbor staging and Cox regression ECOG as well as sex, the MFPT model includes all these variables. On the one hand, this means that each modelling-based score re-captures at least three of five IPI_{Shipp} factors where extranodal manifestations seem negligible when also considering early response to therapy. On the other hand, it indicates that the IPI-based scores lack possibly existing prognostic value of sex and histological diagnosis. Nevertheless, this is only the behaviour with the PETAL data and again needs to be confirmed in an independent sample. The direction of the effects of the selected variables seem plausible from a clinical point of view and their magnitudes appear similar across all three prognostic models. Note that in the logistic regression and MFPT approach risk scores, Ann Arbor stages III and IV do not seem to produce subgroups with different risk which is in line with the dichotomisation in the IPI_{Shipp} (I/II vs. III/IV). The protective effect of SUV at baseline PET/CT looks odd at first sight but can be explained by the fact that higher tumour burden at baseline allows on absolute scale for larger reduction of SUV during the first two cycles of (R-)CHOP that is represented by the effect of SUV at interim PET/CT. Speaking of practicality, the modelling-based prognostic risk scores developed in this thesis contain six, seven, and eight variables in their respective final models which seems to be borderline in terms of complexity. Still, one can imagine some optimisations to further reduce the number of variables. For example, the entire PETAL data set is considered here although the diagnoses it covers are rather heterogeneous and there definitely are differences in overall survival between B-cell and T-cell lymphomas (Pfreundschuh et al., 2008; Ellin et al., 2014). Possibly, such differences may also exist in the set of prognostic variables or the magnitude of their effects. Thus, it may be worthwhile to investigate B- and T-cells separately or restrict the analysis to the DLBCL subgroup that is by far the most common, and thus, most dominant entity of aggressive NHL. DLBCL, however, are

themselves a heterogeneous group of lymphomas. Therefore, inclusion of or stratifying by, for example, the cell of origin as obtained by gene expression profiling can be a reasonable option with germinal centered B-cell and activated B-cell molecular subtypes being the two major categories (Alizadeh et al., 2000) that also received attention in the most recent revision of the WHO classification of lymphoid neoplasms (Swerdlow et al., 2016). When thinking of simplifications of the prognostic risk scores developed here, one should also discuss the set of candidate variables used. With respect to early interim PET/CT response, the absolute and relative differences between SUV at baseline PET/CT and interim PET/CT appear to be reasonable choices instead of including the two variables separately. Extending the set by, e.g., genetic factors (Shipp et al., 2002), metabolic tumour volume (Schmitz et al., 2020), or the Deauville score (Meignan et al., 2009) may lead to completely different final models with a possibly smaller number of variables. For example, certain genes or tumour volume may be surrogates for some of the predictors identified in this thesis, and thus, may reduce the number of variables in the final model. Alternatively using the Deauville criteria seems appealing as it would enlarge the bad prognosis group. However, a post-hoc analysis to the PETAL trial showed that this is due to a high false-positive rate and consequently lacks the desired discrimination performance (Dührsen et al., 2018, Supplemental Table A3).

Cross-validation issues Reconsider validation sample two where the MFPT approach selects a time-dependent function for the weight of age in the training data set. The resulting poor performance of the MFPT-based score compared to the logistic regression and Cox regression scores may have two possible explanations. First, it may be due to overfitting with the model-building process yielding a false-positive decision for a time-dependent term that is not present in the test data set and may not be present in the total population. Note that the associated p-value in favour of considering time-dependency of the age effect is only barely below the 5% alpha level and would not have led to a rejection of proportionality if a stricter criterion had been applied. Second, it may relate to the elicitation of one single number from the time-dependent function as described in Section 2.3. The second explanation reflects the difficulty to consider the time part of the prediction given the fact that the survival time of a newly diagnosed patient will always be unknown. It appears that the weighting approach used in this thesis (weighting by the relative frequency of follow-up time in the entire PETAL sample) is not the optimal choice. Alternatives include naively taking the average estimate over the time interval between interim PET/CT and maximum follow-up time

or taking the weight associated with mean or median survival time. The latter may also be applied to the predicted mean or median survival time of a patient with the same characteristics as the new patient in question. Also, another possibility could be to use multiple imputation for follow-up time and the censoring indicator and to take the median of a large number of imputations. Both approaches would, however, require combining the new patient's data with individual patient data from the risk score development data set. In general, as long as there is no recommendation on how to incorporate a time-dependent effect into a prognostic risk score, using an MFP model that only addresses non-linearity seems to be more robust and safe than the MFPT approach.

A further aspect to discuss is the uniformly bad performance in validation sample number four. Performance is poor for all five investigated risk scores and further investigation shows surprising and implausible behaviour of the variable Ann Arbor staging in this validation sample's test data set. That is, in terms of hazard ratios, patients with stage I are more likely to die early than patients with any other Ann Arbor stage. Also, an LDH value above the upper reference limit seems to be associated with a prolonged survival time in the test data set. Both observations are counter-intuitive and this behaviour is not observed in the respective training data set. This explains why the risk is not modelled satisfactorily and it may indicate that the size of the test data set is too small. However, the integer-based scores IPI_{Shipp} and IPI_{iPET} surprisingly seem to suffer the most from that issue — even experiencing a negative Youden index for their respective optimal cut-off point. This may be explained by the fact that their initial discrimination performance in the entire PETAL data set is already only slightly higher than that of an unbiased coin.

A closer look from statistical perspective This thesis touches several statistical issues that are crucial in the development process of any prognostic risk model:

The choice of overall survival as endpoint to develop the prognostic risk scores is made in opposition to time to treatment failure being the primary endpoint in the PETAL trial. However, overall survival is by all means patient-relevant and does not mix up efficacy and safety aspects like time to treatment failure (McKee et al., 2010). An issue with overall survival is that it can be confounded by subsequent treatments like radiation therapy or intensified chemotherapy (McKee et al., 2010). Also, competing risks can bias the results obtained by standard time-to-event methodology when death from any cause is used (Andersen et al., 2012). One may want to think about surrogate

endpoints for overall survival such as disease-free survival, progression-free survival, and time to progression, and they also were defined as secondary endpoints in the PETAL trial. But whether they predict overall survival in aggressive NHL patients is still unclear although Lee et al. (2011) see first hints for a correlation between 3-year progression-free survival and 5-year overall survival in their meta-analysis with individualised patient data. Also, as an argument against progression-free survival and time to progression, Ciani et al. (2014) argue that these are irrelevant endpoints for curative treatments as progression cannot occur after complete remission — however, relapse can.

Regarding the missing data mechanism, there does not seem to be an obvious reason for the behaviour that patients with higher maximum SUV at interim PET/CT have higher odds of having missing data. Thus, it may just be due to chance given the small sample size. Maximum SUV at interim PET/CT was assessed by the specialist in nuclear medicine so that a lack of documentation of baseline covariates should not have influenced the maximum SUV at interim PET/CT value. Also, the data originated from a randomised clinical trial that provides almost complete data regarding the variables in the candidate set. Given that, there is no further evidence to challenge the MCAR assumption. Therefore, handling of missing data is regarded as a negligible concern in this thesis and complete case analysis seems justifiable. In most cases, however, prognostic risk models are developed on data from observational studies or from registry data where missing information may be more common resulting in a mechanism different from MCAR. Multiple imputation using Rubin's rules (Rubin, 1987) may be applied under these circumstances. With respect to model development, one of the major questions then is how to combine information from imputed data sets with different sets of variables being selected; see Wood et al. (2008) for a review of methods that allow for such combination. For example, one method they consider is successively applying Rubin's rules to each step of backward elimination although they find it to be rather computer-intensive. Already for "simple" MFP models the complexity even increases with not only the necessity to make decisions on the selection of variables but also with respect to their functional form. This specific problem is considered by Morris et al. (2015); nevertheless, an application to the much more complex MFPT approach that is used in this thesis is not yet available neither in the literature nor in statistical software and seems questionable in general.

Speaking of variable selection, there still is much debate on whether to use it at all in statistical model-building (e.g., Altman, 2000). Some advocate its use and backward elimination has been shown to be the most reliable of the methods implemented in

standard statistical software (Royston and Sauerbrei, 2008). But, the decision criterion to remove a variable from the model either remains unclear or differs by the intended use of the model (Heinze and Dunkler, 2017). Others prefer to generally avoid model selection techniques and to instead develop a model with a pre-specified set of the clinically most plausible predictor variables only (e.g., Steyerberg, 2009). Among other issues, critics especially argue that variable selection may lead to inflation of type I error with incorrectly selecting variables without an actual effect and also to biased effect estimates in the final model (see, e.g., Harrell, 2015). While shrinking techniques like the "lasso" (least absolute shrinkage and selection operator; Tibshirani, 1996) may be used to counter the bias issue, bootstrapping may help with the multiple comparisons concern in either assessing robustness of the selected variables (Altman and Andersen, 1989) or going further in including only candidate variables in the final model that exceed a certain inclusion frequency in the bootstrap samples (Sauerbrei and Schumacher, 1992). Nevertheless, such frequency threshold for the bootstrap will always be arbitrary.

With regard to local discrimination performance, the primary diagnostic measure is the closest-to-(0,1) corner criterion while Youden index and concordance probability are reported as sensitivity analyses. The choice of the closest-to-(0,1) corner criterion is made in accordance with Rota and Antolini (2014) who find it to be the least biased among the three measures. Besides the robustness assessment, also the more meaningful interpretation of Youden index and concordance probability motivate their consideration, that is, their intention is to jointly minimise the false positive and the false negative rate. Note that in the analyses of this thesis the choice of the criterion does not seem to play a crucial role as for all risk scores but the one obtained by Cox regression modelling the three criteria result in the same cut-off points. In general, data-driven estimation of such a cut-off point has been much criticised due to the over-optimism it may introduce (see, e.g., Leeflang et al., 2008). However, especially in the PETAL trial the aim was to discriminate patients responding satisfactorily to the standard (R-)CHOP therapy regimen from those more or less resistant to that regimen to find better treatment options for the latter group.

An interesting point is that discrimination and calibration performance do not improve remarkably when using Cox regression instead of logistic regression for risk score development. With the entire data set, logistic regression and Cox regression are on almost the same level; and in the cross-validated samples, logistic regression appears to provide the more robust results. Thus, it seems that most of the information contained in the time-to-event variable overall survival is already explained by the binary

2-year survival variable. An explanation from the clinical side is that in aggressive NHL tumour-related fatalities are more likely to occur over the course of the first two years after diagnosis rather than thereafter (Cunningham et al., 2013) such that this time point marks a milestone in aggressive NHL therapy. Regarding the comparison of logistic regression and Cox regression, the influence of later deaths that are only addressed by the time-to-event variable is presumably negligible. Consequently, two years after interim PET/CT defines the binary outcome variable in logistic regression in this thesis and is as well used for t_{short} in the MFPT approach and also for the time point of interest τ for the time-dependent ROC analysis. From statistical perspective, the choice of τ for the approach of Antolini and Valsecchi (2012) may also play a role and appears to mimic the scenario of a traditional ROC analysis with the binary 2-year survival variable. This explanation underlines the weakness of the time-dependent ROC analysis used in this thesis that a time point of interest needs to be specified. However, this also applies to the more popular method of Heagerty et al. (2000) and any other method aiming at t-year survival. Alternative approaches that do not suffer from this weakness are usually based on Harrell's well-known C index (Harrell et al., 1982; Uno et al., 2011). But note that Harrell's C utilises only part of the data as it ignores pairs of observations with the shorter time being censored. In general, there is no gold standard in time-dependent ROC analysis (Wolf et al., 2011) and comparison across studies is presumably difficult. To clarify the actual gain of time-dependent over standard ROC analysis in the aggressive NHL example, the PETAL data may be used for an investigation that compares time-dependent ROC analyses at different choices of τ with standard ROC analyses using the respective binary variables.

Drawbacks and limitations In terms of possible violations of the proportional hazards assumption, a drawback of this thesis is the lack of parametric accelerated failure time models (Collett, 2014, Chapter 6). As they do not make an assumption on the proportionality of hazards, they are a worthwhile alternative to Cox regression and the MFPT approach. Additionally, in contrast to Cox regression and its derivatives they allow for an interpretation in terms of prolonging survival time and for the prediction of a new patient's individual survival time (Swindell, 2009). They have not been considered in this thesis but it seems appealing to pursue this effort in a follow-up investigation. However, it should be noted that in being parametric models the relaxation of the proportional hazards assumption comes at the price of survival time following a certain parametric probability distribution like, e.g., the Weibull distribution which can either

be pre-specified or elicited via model selection methods. Yet another class of approaches of possible interest are machine learning techniques (Obermeyer and Emanuel, 2016) like, e.g., artificial and deep neural networks that have received great attention in recent years. It would be very tempting to see how their discrimination and calibration performance behaves as compared to the more traditional statistical methods utilised in this thesis.

Another limitation of this thesis is that bootstrap methods would be a more reliable tool than internal 8-fold cross-validation to assess robustness of the results using the data at hand (Efron, 1983; Efron and Tibshirani, 1997). First of all, the decision for the cross-validation to be 8-fold is rather arbitrary. One justification can be that with the data set divided into eight parts and 862 patients in the data set each of the sub-data sets has a reasonable size of more than one hundred patients. The bootstrap would require constructing a fair amount of resamples with replacement from the observed data — usually hundreds or thousands. For the analysis, however, this would include repeating the whole model-building process as well as the cut-off point determination and assessment for each bootstrap iteration instead of for eight validation samples. This was not feasible in the limited time frame of this thesis for the MFPT approach due to the complexity of the algorithm and hence computational reasons. If the bootstrap were used to obtain one single final prognostic risk model to be applied in clinical practice, again the question would arise how to combine information from the variable selection process of a large number of bootstrap iterations. Techniques from the methods described by Wood et al. (2008) and Morris et al. (2015) that allow for combining multiple imputation and variable selection may be used but theoretical justification for applying Rubin's rules to bootstrapped data sets remains unclear.

Outlook As a general outlook, an external validation study is required to confirm the modelling-based scores' features in an independent patient population before they can be used in clinical practice. Such a study would not only offer external validation but would also allow for re-calibration of the proposed models in refining the selection and weighting of prognostic variables. Once a final prognostic risk score for aggressive NHL patients under (R-)CHOP therapy is obtained, it may be presented as an online tool or a smartphone application. Apart from an external validation study, the PETAL data themselves offer the chance to reconsider methodological issues. Simulation studies may be carried out to substantiate decisions on statistical actions and approaches. The discussions on missing data, proportional hazards assessment, variable selection, time-dependent ROC methods, and model-building approaches in general show that

although the series on prognostic research (Steyerberg et al., 2013; Moons et al., 2009) provide general frameworks for the development of prognostic models, it lacks on recommendations on their actual execution — that is, which methods should be used and how should they be used. The number of available and suitable methods is manifold and the researcher may decide on these topics to the best of his or her knowledge and personal preference, and thus, different researchers may produce different results, models, and scores — possibly undermining their credibility. This thesis addresses the need for more comparative work from the perspective of an applied statistician (Boulesteix et al., 2018) but more research and guidance is needed here. The STRATOS (STRengthening Analytical Thinking in Observational Studies; Sauerbrei et al., 2014) initiative may be a good start as they only recently began their work on best practices in the analysis of observational studies and their topic groups among other issues tackle the major part of critical aspects identified in this thesis such as missing data handling, functional forms for continuous variables, diagnostic test and prediction model evaluation, and survival analysis in general.

5 Summary

Optimisation of treatment decisions in aggressive Non-Hodgkin lymphomas (NHL) is necessarily needed as many patients still do not respond well to standard therapy — a chemotoxic regimen consisting of cyclophosphamide, doxorubicin, vincristine, prednisone, and, optionally, rituximab. Using data from the Positron Emission Tomography–Guided Therapy of Aggressive Non-Hodgkin Lymphomas (PETAL) trial, this thesis introduces three prognostic risk scores built from statistical modelling approaches and compares them with the well-established International Prognostic Index (IPI). The main clinical feature of these scores is that they in contrast to the IPI utilise positron emission tomography imaging to consider early response to therapy as assessed by the metabolic activity of the tumour after the first two cycles of chemotherapy. From statistical point of view, the three modelling approaches logistic regression, Cox regression, and the multivariable fractional polynomial time (MFPT) approach vary in their methodological assumptions. That is, in contrast to the other two, the latter technique can model non-linear relationships between prognostic variables and outcome while it also allows for the modelled effect to be time-dependent. The results of this thesis propose that early response to therapy measured as the relative reduction between pre- and post-chemotherapy standardised uptake value of a radioactive tracer does have an additional prognostic value beyond the IPI and that the modelling-based scores appear to be worth the effort as compared to simple one-point penalty scores like the IPI. With respect to the MFPT approach, the possibility to use non-linear transformations seems to improve discrimination and calibration performance while time-dependent effects appear to play a minor role in the aggressive NHL setting. The prognostic risk score obtained by the MFPT approach reveals good performance in identifying patients with an unfavourable prognosis under standard treatment and clearly outperforms the IPI. Therefore, it may be helpful in guiding decisions on treatment intensification in aggressive NHL. Nevertheless, any prognostic risk score developed in this thesis needs to be validated in an independent patient population before it shall be used in clinical routine.

6 References

1. Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E., Moore T., Hudson J. Jr., Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C., Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Byrd J.C., Botstein D., Brown P.O. and Staudt L.M. (2000): Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511.

2. Altman D.G. (2000): Statistics in medical journals: some recent trends. Stat. Med. 19, 3275–3289.

3. Altman D.G. (2009): Prognostic models: a methodological framework and review of models for breast cancer. Cancer. Invest. 27, 235–243.

4. Altman D.G. and Andersen P.K. (1989): Bootstrap investigation of the stability of a Cox regression model. Stat. Med. 8, 771–783.

5. Altman D.G. and Royston P. (2000): What do we mean by validating a prognostic model? Stat. Med. 19, 453–473.

6. Andersen P.K., Geskus R.B., de Witte T. and Putter H. (2012): Competing risks in epidemiology: possibilities and pitfalls. Int. J. Epidemiol. 41, 861–870.

7. Antolini L., Boracchi P. and Biganzoli E. (2005): A time-dependent discrimination index for survival data. Stat. Med. 24, 3927–3944.

8. Antolini L. and Valsecchi M.G. (2012): Performance of binary markers for censored failure time outcome: nonparametric approach based on proportions. Stat. Med. 31, 1113–1128.

9. Armitage J.O., Gascoyne R.D., Lunning M.A. and Cavalli F. (2017): Non-Hodgkin lymphoma. Lancet 390, 298–310.

10. Austin P.C. and Steyerberg E.W. (2012): Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med. Res. Methodol. 12, 82.

11. Barlow W.E. and Prentice R.L. (1988): Residuals for relative risk regression. Biometrika 75, 65–74.

12. Boulesteix A.L., Binder H., Abrahamowicz M. and Sauerbrei W. for the Simulation Panel of the STRATOS Initiative (2018): On the necessity and design of studies comparing statistical methods. Biom. J. 60, 216–218.

13. Brier G.W. (1950): Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3.

14. Buchholz A. and Sauerbrei W. (2011): Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. Biom. J. 53, 308–331.

15. Carbone P.P., Kaplan H.S., Musshoff K., Smithers D.W. and Tubiana M. (1971): Report of the committee on Hodgkin's disease staging classification. Cancer Res. 31, 1860–1861.

16. Cheson B.D., Fisher R.I., Barrington S.F., Cavalli F., Schwartz L.H., Zucca E. and Lister T.A. (2014): Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. J. Clin. Oncol. 32, 3059.

17. Ciani O., Davis S., Tappenden P., Garside R., Stein K., Cantrell A., Saad E.D., Buyse M. and Taylor R.S. (2014): Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. Int. J. Technol. Assess. Health Care 30, 312–324.

18. Clark T.G., Altman D.G. and De Stavola B.L. (2002): Quantification of the completeness of follow-up. Lancet 359, 1309–1310.

19. Collett D. (2014): Modelling Survival Data in Medical Research. 3rd Ed. New York: Chapman and Hall/CRC.

20. Cox D.R. (1972): Regression models and life-tables. J. Royal Stat. Soc. B 34, 187–202.

21. Cunningham D., Hawkes E.A., Jack A., Qian W., Smith P., Mouncey P., Pocock C., Ardeshna K.M., Radford J.A., McMillan A., Davies J., Turner D., Kruger A., Johnson P., Gambell J. and Linch D. (2013): Rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisolone in patients with newly diagnosed diffuse large B-cell non-Hodgkin lymphoma: a phase 3 comparison of dose intensification with 14-day versus 21-day cycles. Lancet 381, 1817–1826.

22. Du Bois D. and Du Bois E.F. (1916): Clinical calorimetry tenth paper: a formula to estimate the approximate surface area if height and weight be known. Arch. Intern. Med. 17, 863–871.

23. Dührsen U., Hüttmann A., Jöckel K.H. and Müller S. (2009): Positron emission tomography guided therapy of aggressive non-Hodgkin lymphomas — the PETAL trial. Leuk. Lymphoma 50, 1757–1760.

24. Dührsen U., Müller S., Hertenstein B., Thomssen H., Kotzerke J., Mesters R., Berdel W.E., Franzius C., Kroschinsky F., Weckesser M., Kofahl-Krause D., Bengel F.M., Dürig J., Matschke J., Schmitz C., Poeppel T., Ose C., Brinkmann M., La Rosée P., Freesmeyer M., Hertel A., Höffkes H.G., Behringer D., Prange-Krex G., Wilop S., Krohn T., Holzinger J., Griesshammer M., Giagounidis A., Raghavachar A., Maschmeyer G., Brink I., Bernhard H., Haberkorn U., Gaska T., Kurch L., van Assema D.M., Klapper W., Hoelzer D., Geworski L., Jöckel K.H., Scherag A., Bockisch A., Rekowski J. and Hüttmann A. for the PETAL Trial Investigators (2018): Positron emission tomography-guided therapy of aggressive non-Hodgkin lymphomas (PETAL): A multicenter, randomized phase III trial. J. Clin. Oncol. 36, 2024–2034.

25. Efron B. (1983): Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78, 316–331.

26. Efron B. and Tibshirani R. (1997): Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92, 548–560.

27. Ellin F., Landström J., Jerkeman M. and Relander T. (2014): Real world data on prognostic factors and treatment in peripheral T-cell lymphomas: a study from the Swedish Lymphoma Registry. Blood 124, 1570–1577.

28. Fisher R.I., Gaynor E.R., Dahlberg S., Oken M.M., Grogan T.M., Mize E.M., Glick J.H., Coltman Jr. C.A. and Miller T.P. (1993): Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma. N. Engl. J. Med. 328, 1002–1006.

29. Geisser S. (1975): The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70, 320–328.

30. Gerds T.A., Cai T. and Schumacher M. (2008): The performance of risk prediction models. Biom. J. 50, 457–479.

31. Graf E., Schmoor C., Sauerbrei W. and Schumacher M. (1999): Assessment and comparison of prognostic classification schemes for survival data. Stat. Med. 18, 2529–2545.

32. Gray R.J. (1992): Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J. Am. Stat. Assoc. 87, 942–951.

33. Hanley J.A. and McNeil B.J. (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

34. Harrell F.E. Jr. (2015): Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd Ed. Cham et al.: Springer International Publishing.

35. Harrell F.E. Jr., Califf R.M., Pryor D.B., Lee K.L. and Rosati R.A. (1982): Evaluating the yield of medical tests. JAMA 247, 2543–2546.

36. Harrell F.E. Jr., Lee K.L. and Pollock B.G. (1988): Regression models in clinical studies: determining relationships between predictors and response. J. Natl. Cancer Inst. 80, 1198–1202.

37. Heagerty P.J., Lumley T. and Pepe M.S. (2000): Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56, 337–344.

38. Heinze G. and Dunkler D. (2017): Five myths about variable selection. Transpl. Int. 30, 6–10.

39. Hermans J., Krol A.D., van Groningen K., Kluin P.M., Kluin-Nelemans J.C., Kramer M.H., Noordijk E.M., Ong F. and Wijermans P.W. (1995): International Prognostic Index for aggressive non-Hodgkin's lymphoma is valid for all malignancy grades. Blood 86, 1460–1463.

40. Hoelzer D., Walewski J., Döhner H., Viardot A., Hiddemann W., Spiekermann K., Serve H., Dührsen U., Hüttmann A., Thiel E., Dengler J., Kneba M., Schaich M., Schmidt-Wolf I.G., Beck J., Hertenstein B., Reichle A., Domanska-Czyz K., Fietkau R., Horst H.A., Rieder H., Schwartz S., Burmeister T. and Gökbuget N. for the German Multicenter Study Group for Adult Acute Lymphoblastic Leukemia (2014): Improved outcome of adult Burkitt lymphoma/leukemia with rituximab and chemotherapy: report of a large prospective multicenter trial. Blood 124, 3870–3879.

41. Hosmer D.W., Lemeshow S. and Sturdivant R.X. (2013): Applied Logistic Regression. 3rd Ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

42. Hüttmann A., Rekowski J., Müller S., Hertenstein B., Franzius C., Mesters R., Weckesser M., Kroschinsky F., Kotzerke J., Ganser A., Bengel F.M., La Rosée P., Freesmeyer M., Höffkes H.G., Hertel A., Behringer D., Prange-Krex G., Griesshammer M., Holzinger J., Wilop S., Krohn T., Raghavachar A., Maschmeyer G., Brink I., Schroers R., Gaska T., Bernhard H., Giagounidis A., Schütte J., Dienst A., Hautzel H., Naumann R., Klein A., Hahn D., Pöpperl G., Grube M., Marienhagen J., Schwarzer A., Kurch L., Höhler T., Steiniger H., Nückel H., Südhoff T., Römer W., Brinkmann M., Ose C., Alashkar F., Schmitz C., Dürig J., Hoelzer D., Jöckel K.H., Klapper W. and Dührsen U. (2019): Six versus eight doses of rituximab in patients with aggressive B cell lymphoma receiving six cycles of CHOP: results from the "Positron Emission Tomography-Guided Therapy of Aggressive Non-Hodgkin Lymphomas" (PETAL) trial. Ann. Hematol. 98, 897–907.

43. Ibrahim J.G., Chen M.H. and Sinha D. (2001): Bayesian Survival Analysis. New York et al.: Springer Science+Business Media.

44. Kaplan E.L. and Meier P. (1958): Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53, 457–481.

45. Lachin J.M. (1999): Worst-rank score analysis with informatively missing observations in clinical trials. Control. Clin. Trials 20, 408–422.

46. Lee L., Wang L. and Crump M. (2011): Identification of potential surrogate end points in randomized clinical trials of aggressive and indolent non-Hodgkin's lymphoma: correlation of complete response, time-to-event and overall survival end points. Ann. Oncol. 22, 1392–1403.

47. Leeflang M.M., Moons K.G., Reitsma J.B. and Zwinderman A.H. (2008): Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin. Chem. 54, 729–737.

48. Lin C., Itti E., Haioun C., Petegnief Y., Luciani A., Dupuis J., Paone G., Talbot J.N., Rahmouni A. and Meignan M. (2007): Early ^{18}F-FDG PET for prediction of prognosis in patients with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis. J. Nucl. Med. 48, 1626–1632.

49. Lin D.Y., Wei L.J. and Ying Z. (1993): Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 80, 557–572.

50. Lister T.A., Crowther D., Sutcliffe S.B., Glatstein E., Canellos G.P., Young R.C., Rosenberg S.A., Coltman C.A. and Tubiana M. (1989): Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. J. Clin. Oncol. 7, 1630–1636.

51. Little R.J. and Rubin D.B. (2002): Statistical Analysis with Missing Data. 2nd Ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

52. Liu X. (2012): Classification accuracy and cut point selection. Stat. Med. 31, 2676–2686.

53. Marcus R., Eric P. and Gabriel K.R. (1976): On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63, 655–660.

54. McKee A.E., Farrell A.T., Pazdur R. and Woodcock J. (2010): The role of the US Food and Drug Administration review process: clinical trial endpoints in oncology. Oncologist 15, Suppl. 1, 13–18.

55. Meignan M., Gallamini A. and Haioun C. (2009): Report on the first international workshop on interim-PET scan in lymphoma. Leuk. Lymphoma 50, 1257–1260.

56. Michallet A.S. and Coiffier B. (2009): Recent developments in the treatment of aggressive non-Hodgkin lymphoma. Blood Rev. 23, 11–23.

57. Mogensen U.B., Ishwaran H. and Gerds T.A. (2012): Evaluating random forests for survival analysis using prediction error curves. J. Stat. Softw. 50, 1–23.

58. Moons K.G., Kengne A.P., Woodward M., Royston P., Vergouwe Y., Altman D.G. and Grobbee D.E. (2012): Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 98, 683–690.

59. Moons K.G., Royston P., Vergouwe Y., Grobbee D.E. and Altman D.G. (2009): Prognosis and prognostic research: what, why, and how? BMJ 338, b375.

60. Morris T.P., White I.R., Carpenter J.R., Stanworth S.J. and Royston P. (2015): Combining fractional polynomial model building with multiple imputation. Stat. Med. 34, 3298–3317.

61. Obermeyer Z. and Emanuel E.J. (2016): Predicting the future — big data, machine learning, and clinical medicine. N. Engl. J. Med. 375, 1216–1219.

62. Oken M.M., Creech R.H., Tormey D.C., Horton J., Davis T.E., McFadden E.T. and Carbone P.P. (1982): Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am. J. Clin. Oncol. 5, 649–656.

63. Park J.H., Yoon D.H., Kim D.Y., Kim S., Seo S., Jeong Y., Lee S.W., Park C.S., Huh J. and Suh C. (2014): The highest prognostic impact of LDH among International Prognostic Indices (IPIs): an explorative study of five IPI factors among patients with DLBCL in the era of rituximab. Ann. Hematol. 93, 1755–1764.

64. Pencina M.J., D'Agostino R.B. and Vasan R.S. (2010): Statistical methods for assessment of added usefulness of new biomarkers. Clin. Chem. Lab. Med. 48, 1703–1711.

65. Perkins N.J. and Schisterman E.F. (2006): The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am. J. Epidemiol. 163, 670–675.

66. Pfreundschuh M., Schubert J., Ziepert M., Schmits R., Mohren M., Lengfelder E., Reiser M., Nickenig C., Clemens M., Peter N., Bokemeyer C., Eimermacher H., Ho A., Hoffmann M., Mertelsmann R., Trümper L., Balleisen L., Liersch R., Metzner B., Hartmann F., Glass B., Poeschel V., Schmitz N., Ruebe C., Feller A.C. and Loeffler M. for the German High-Grade Non-Hodgkin Lymphoma Study Group (DSHNHL) (2008): Six versus eight cycles of bi-weekly CHOP-14 with or without rituximab in elderly patients with aggressive CD20+ B-cell lymphomas: a randomised controlled trial (RICOVER-60). Lancet Oncol. 9, 105–116.

67. Rota M. and Antolini L. (2014): Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers. Comput. Stat. Data Anal. 69, 1–14.

68. Rota M., Antolini L. and Valsecchi M.G. (2015): Optimal cut-point definition in biomarkers: the case of censored failure time outcome. BMC Med. Res. Methodol. 15, 24.

69. Royston P. and Altman D.G. (1994): Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. J. Royal Stat. Soc. C 43, 429–453.

70. Royston P., Altman D.G. and Sauerbrei W. (2006): Dichotomizing continuous predictors in multiple regression: a bad idea. Stat. Med. 25, 127–141.

71. Royston P. and Sauerbrei W. (2008): Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Chichester, UK: John Wiley & Sons, Ltd.

72. Rubin D.B. (1987): Multiple Imputation for Nonresponse in Surveys. New York et al.: John Wiley & Sons, Inc.

73. Sauerbrei W., Abrahamowicz M., Altman D.G., le Cessie S. and Carpenter C. for the STRATOS initiative (2014): STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Stat. Med. 33, 5413–5432.

74. Sauerbrei W. and Royston P. (1999): Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J. Royal Stat. Soc. A 162, 71–94.

75. Sauerbrei W., Royston P. and Look M. (2007): A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biom. J. 49, 453–473.

76. Sauerbrei W. and Schumacher M. (1992): A bootstrap resampling procedure for model building: application to the Cox regression model. Stat. Med. 11, 2093–2109.

77. Schemper M. and Smith T.L. (1996): A note on quantifying follow-up in studies of failure time. Control. Clin. Trials 17, 343–346.

78. Schmitz C., Hüttmann A., Müller S., Hanoun M., Boellaard R., Brinkmann M., Jöckel K.H., Dührsen U. and Rekowski J. (2019): Dynamic risk assessment based on positron emission tomography scanning in diffuse large B-cell lymphoma. Manuscript submitted for publication.

79. Schmitz C., Hüttmann A., Müller S., Hanoun M., Boellaard R., Brinkmann M., Jöckel K.H., Dührsen U. and Rekowski J. (2020): Dynamic risk assessment based on positron emission tomography scanning in diffuse large B-cell lymphoma: Post-hoc analysis from the PETAL trial. Eur. J. Cancer 124, 25–36.

80. Schoenfeld D. (1982): Partial residuals for the proportional hazards regression model. Biometrika 69, 239–241.

81. Schulz K.F., Altman D.G. and Moher D. (2010): CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC Med. 8, 18.

82. Sehn L.H., Berry B., Chhanabhai M., Fitzgerald C., Gill K., Hoskins P., Klasa R., Savage K.J., Shenkier T., Sutherland J., Gascoyne R.D. and Connors J.M. (2007): The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. Blood 109, 1857–1861.

83. Shankland K.R., Armitage J.O. and Hancock B.W. (2012): Non-Hodgkin lymphoma. Lancet 380, 848–857.

84. Shipp M.A., Harrington D.P., Anderson J.R., Armitage J.O., Bonadonna G., Brittinger G., Cabanillas F., Canellos G.P., Coiffier B., Connors J.M., Cowan R.A., Crowther D., Dahlberg S., Engelhard M., Fisher R.I., Gisselbrecht C., Horning S.J., Lepage E., Lister T.A., Meerwaldt J.H., Montserrat E., Nissen N.I., Oken M.M., Peterson B.A., Tondini C., Velasquez W.A. and Yeap B.Y. (1993): A predictive model for aggressive non-Hodgkin's lymphoma. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. N. Engl. J. Med. 329, 987–994.

85. Shipp M.A., Ross K.N., Tamayo P., Weng A.P., Kutok J.L., Aguiar R.C., Gaasenbeek M., Angelo M., Reich M., Pinkus G.S., Ray T.S., Koval M.A., Last K.W., Norton A., Lister T.A., Mesirov J., Neuberg D.S., Lander E.S., Aster J.C. and Golub T.R. (2002): Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 8, 68–74.

86. Steyerberg E.W. (2009): Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York et al.: Springer Science+Business Media.

87. Steyerberg E.W., Moons K.G., van der Windt D.A., Hayden J.A., Perel P., Schroter S., Riley R.D., Hemingway H. and Altman D.G. for the PROGRESS Group (2013): Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 10, e1001381.

88. Stone M. (1974): Cross-validatory choice and assessment of statistical predictions. J. Royal Stat. Soc. B 36, 111–133.

89. Swerdlow S.H., Campo E., Pileri S.A., Harris N.L., Stein H., Siebert R., Advani R., Ghielmini M., Salles G.A., Zelenetz A.D. and Jaffe E.S. (2016): The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood 127, 2375–2390.

90. Swindell W.R. (2009): Accelerated failure time models provide a useful statistical framework for aging research. Exp. Gerontol. 44, 190–200.

91. Tibshirani R. (1996): Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. B 58, 267–288.

92. Uno H., Cai T., Pencina M.J., D'Agostino R.B. and Wei L.J. (2011): On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat. Med. 30, 1105–1117.

93. Van Houwelingen H.C. (2000): Validation, calibration, revision and combination of prognostic survival models. Stat. Med. 19, 3401–3415.

94. Wahl R.L., Jacene H., Kasamon Y. and Lodge M.A. (2009): From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J. Nucl. Med. 50, Suppl. 1, 122S–150S.

95. Wolf P., Schmidt G. and Ulm K. (2011): The use of ROC for defining the validity of the prognostic index in censored data. Stat. Probabil. Lett. 81, 783–791.

96. Wood A.M., White I.R. and Royston P. (2008): How should variable selection be performed with multiply imputed data? Stat. Med. 27, 3227–3246.

97. Youden W.J. (1950): Index for rating diagnostic tests. Cancer 3, 32–35.

98. Zhou Z., Sehn L.H., Rademaker A.W., Gordon L.I., LaCasce A.S., Crosby-Thompson A., Vanderplas A., Zelenetz A.D., Abel G.A., Rodriguez M.A., Nademanee A., Kaminski M.S., Czuczman M.S., Millenson M., Niland J., Gascoyne R.D., Connors J.M., Friedberg J.W. and Winter J.N. (2014): An enhanced International Prognostic Index (NCCN-IPI) for patients with diffuse large B-cell lymphoma treated in the rituximab era. Blood 123, 837–842.

99. Ziepert M., Hasenclever D., Kuhnt E., Glass B., Schmitz N., Pfreundschuh M. and Loeffler M. (2010): Standard International prognostic index remains a valid predictor of outcome for patients with aggressive CD20+ B-cell lymphoma in the rituximab era. J. Clin. Oncol. 28, 2373–2380.

Appendix

Table 8: Odds ratios obtained by a logistic regression model with an indicator variable for missing data as dependent variable and the respective baseline covariate as independent variable.

Baseline covariate	No. of observations with available data in variable of interest	No. of observations with missing data in remaining baseline covariates	Odds ratio [95% confidence interval]
Age (per ten years)	862	15	0.91 [0.65; 1.27]
ECOG	858	11	
1 vs. 0			1.04 [0.26; 4.18]
2 vs. 0			0.91 [0.65; 15.4]
3 vs. 0			11.6 [1.98; 67.6]
AnnArb	859	12	
II vs. I			0.77 [0.11; 5.55]
III vs. I			1.52 [0.26; 8.41]
IV vs. I			0.96 [0.18; 5.33]
LDH	858	11	2.15 [0.57; 8.16]
Sex (male vs. female)	862	15	1.50 [0.54; 4.16]
SUVbase (per 10 units)	858	11	0.93 [0.53; 1.64]
SUVint (per 10 units)	857	10	2.22 [1.20; 4.08]
Diag	862	15	
OthB vs. DLBCL			0.68 [0.08; 5.60]
T vs. DLBCL			2.32 [0.47; 11.4]
Other vs. DLBCL			9.56 [2.92; 31.3]
CD20	862	15	0.69 [0.15; 3.10]
BSympt	859	12	1.62 [0.51; 5.16]
Resect	858	11	n/a
MajOrg	861	14	0.91 [0.28; 2.94]
BSA	857	10	2.58 [0.13; 51.4]

Table 9: Likelihood ratio test for ECOG and Ann Arbor staging to assess whether the variables can be included in the set of candidate variables in continuous form. The models refer to Cox regression models with overall survival as outcome variable and the prognostic variable either addressed in continuous or categorical form.

Explanatory variable	Log-likelihood	χ^2 test statistic	Degrees of freedom	p-value
ECOG (continuous)	-1170.2	0.5608	2	0.7555
ECOG (categorical)	-1169.9			
Ann Arbor (continuous)	-1178.0	10.271	2	0.0059
Ann Arbor (categorical)	-1172.8			

Table 10: Concordance between the five risk scores regarding their prognosis based on the closest-to-(0,1) corner criterion; concordance between two scores is evaluated as their fraction of agreement.

	IPI_{iPET}	LogReg	CoxReg	MFPT
IPI_{Shipp}	0.948	0.744	0.695	0.721
IPI_{iPET}		0.763	0.716	0.742
LogReg			0.819	0.890
CoxReg				0.871

Table 11: Cross-tabulation of the two-year (τ) survival status and the respective risk score prognosis according to the closest-to-(0,1) corner criterion.

		Deceased until τ	Alive at τ	Censored before τ
IPI_{Shipp}	Bad	79	201	27
	Good	49	459	46
IPI_{iPET}	Bad	88	227	32
	Good	37	429	40
LogReg	Bad	89	144	28
	Good	36	512	44
CoxReg	Bad	99	224	28
	Good	26	431	44
MFPT	Bad	89	150	22
	Good	36	505	50

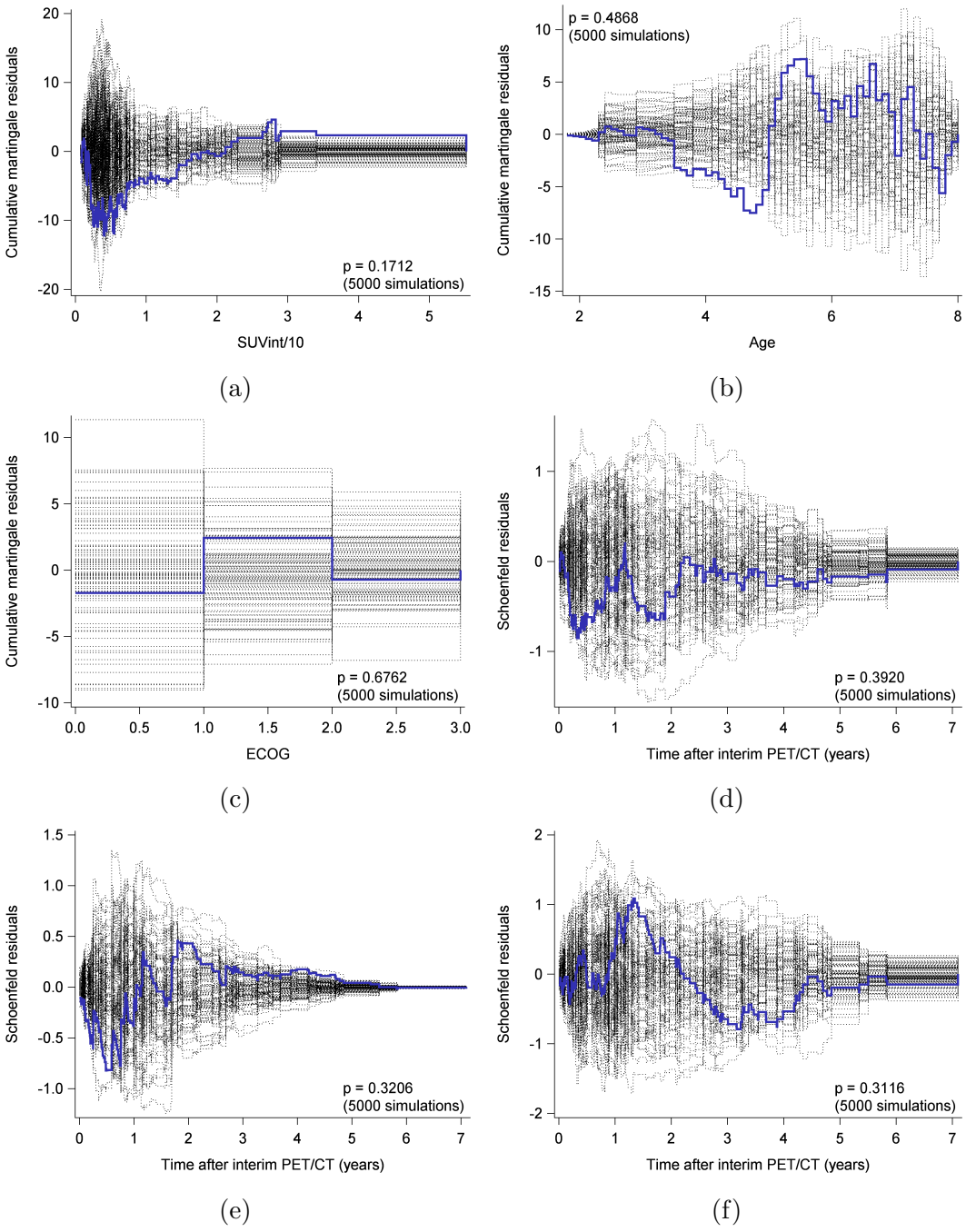


Figure 22: Checking the functional form for (a) maximum SUV at baseline PET/CT, (b) age, and (c) ECOG. Checking the proportional hazards assumption for (d) maximum SUV at baseline PET/CT, (e) maximum SUV at interim PET/CT, and (f) lymphomatous involvement in major organs. The first fifty of $5,000$ simulated patterns are represented as dotted lines; the actually observed path is displayed as solid blue line.

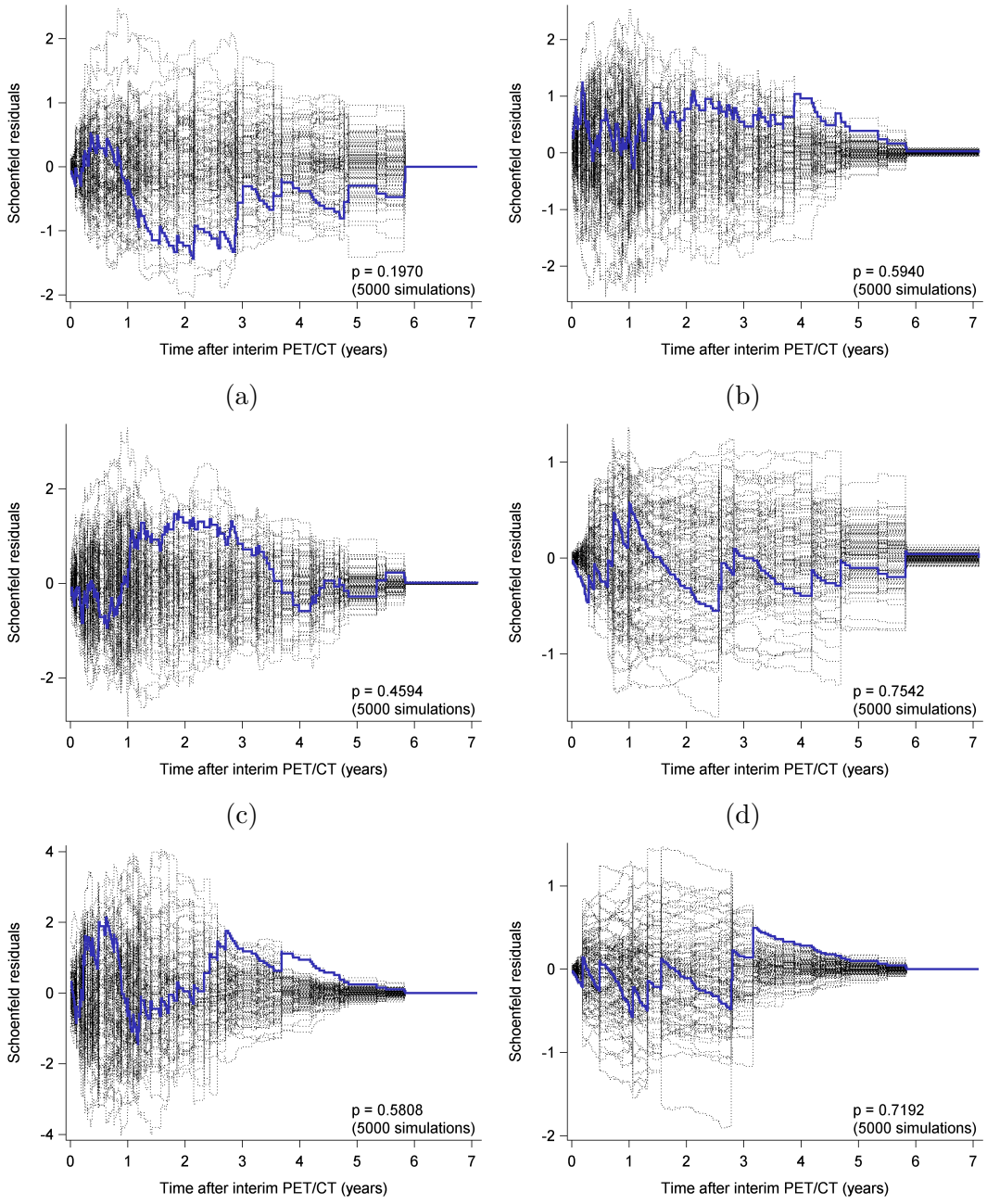


Figure 23: Checking the proportional hazards assumption for (a) Ann Arbor stage II versus I, (b) Ann Arbor stage III versus I, (c) Ann Arbor stage IV versus I, (d) other B-cell lymphomas versus DLBCL, (e) T-cell lymphomas versus DLBCL, and (f) other lymphomas versus DLBCL. The first fifty of $5,000$ simulated patterns are represented as dotted lines; the actually observed path is displayed as solid blue line.

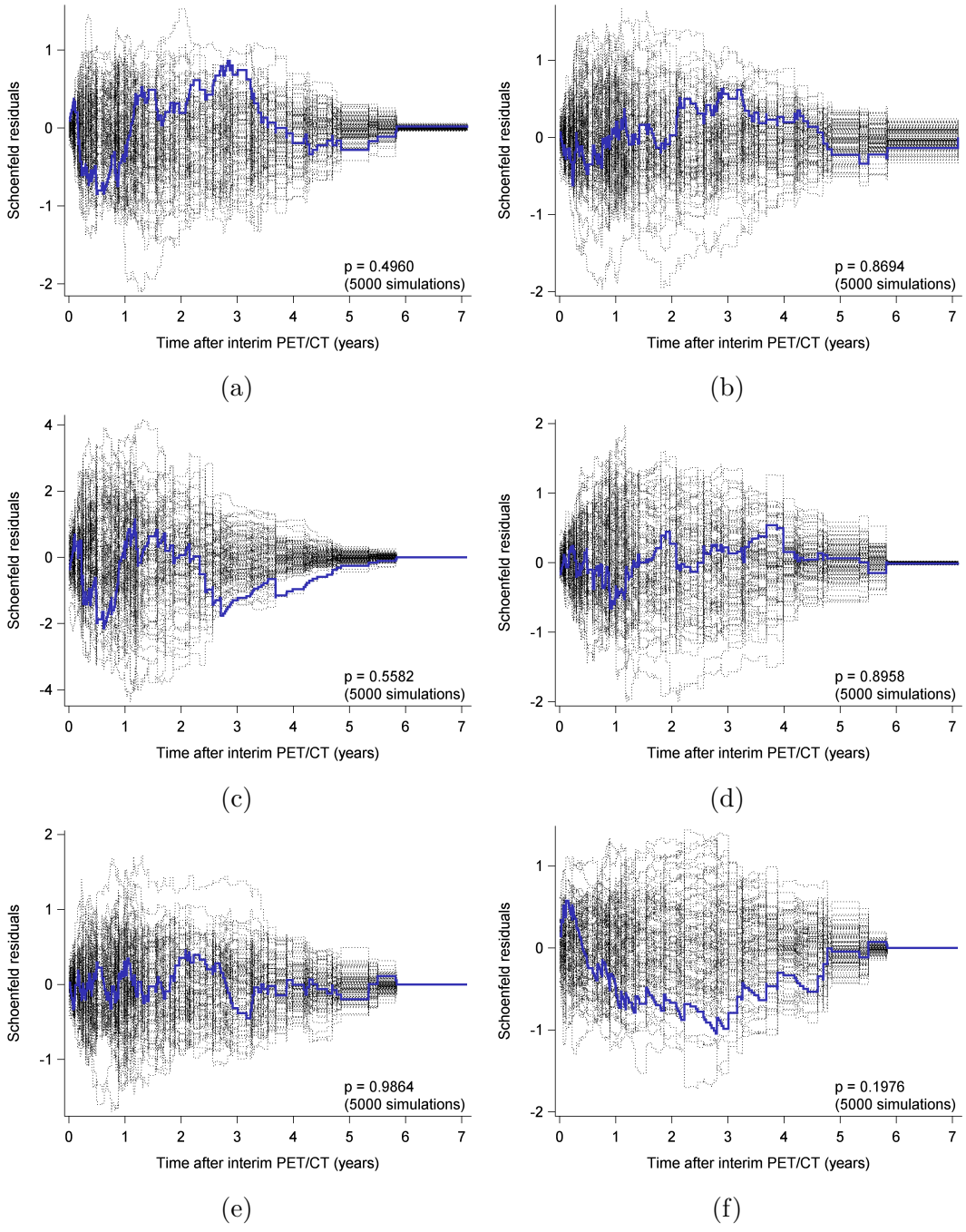


Figure 24: Checking the proportional hazards assumption for (a) LDH, (b) sex, (c) CD20 expression, (d) BSA, (e) B symptoms, and (f) completely resected manifestations. The first fifty of $5,000$ simulated patterns are represented as dotted lines; the actually observed path is displayed as solid blue line.

Table 12: Discrimination and calibration results for the five prognostic risk scores as described and summarised graphically in Section 3.3.

	ED	J	CP	AUC [95% CI]	IBS (reference: 0.151)
IPI_{Shipp}	0.489	0.312	0.429	0.708 [0.655; 0.760]	0.136
IPI_{iPET}	0.460	0.352	0.456	0.741 [0.690; 0.793]	0.131
LogReg	0.368	0.483	0.549	0.802 [0.755; 0.850]	0.118
CoxReg	0.402	0.455	0.522	0.800 [0.752; 0.848]	0.117
MFPT	0.371	0.478	0.545	0.817 [0.771; 0.863]	0.113

Table 13: Clinical characteristics of the PETAL trial intention-to-treat population ($N = 862$) by validation sample four development and test data set. Represented as mean ± standard deviation (median [Q1; Q3]) for continuous variables and percentages for categorical variables.

Baseline covariate	Development data set	Test data set
Age	57.05 ± 15.00	61.07 ± 12.26
	(59 [48; 69])	(64.5 [54; 70])
Sex		
Male	55.7 %	63.2 %
Female	44.3 %	36.8 %
Diag		
DLBCL	70.9 %	72.6 %
OthB	15.2 %	12.3 %
T	8.5 %	10.4 %
Other	5.4 %	4.7 %
SUVbase	20.92 ± 10.65	20.33 ± 12.64
	(19.90 [12.80; 26.80])	(16.23 [11.55; 27.30])
SUVint	4.33 ± 4.09	4.68 ± 4.98
	(3.40 [2.30; 4.80])	(3.30 [2.43; 4.80])
ECOG		
0	47.4 %	39.6 %
1	43.9 %	50.9 %
2	6.7 %	7.5 %
3	2.0 %	1.9 %
AnnArb		
I	17.5 %	17.9 %
II	22.3 %	26.4 %
III	23.6 %	19.8 %
IV	36.6 %	35.8 %
MajOrg	29.7 %	35.8 %
LDH	55.1 %	57.5 %
IPI_{Shipp}		
Low risk	38.7 %	36.8 %
Low-intermediate risk	25.8 %	28.3 %
High-intermediate risk	21.3 %	18.9 %
High risk	14.2 %	16.0 %
CD20	91.5 %	89.6 %
BSympt	30.8 %	29.2 %
Resect	11.7 %	8.5 %
BSA	1.91 ± 0.21	1.91 ± 0.21
	(1.92 [1.77; 2.02])	(1.92 [1.80; 2.00])

Table 14: Hazard ratios (with 95% confidence intervals) for overall survival in the PETAL trial intention-to-treat population ($N = 862$) by validation sample four development and test data set. Hazards ratios are obtained by multivariable Cox regression models with overall survival as outcome variable and the variables of the final models for validation sample four as prognostic variables. Note that with the development data set of validation sample four, logistic regression, Cox regression, and MFPT approach select the same six variables for their respective final models.

Baseline covariate	Development data set	Test data set
Age	1.05 [1.03; 1.06]	1.05 [1.01; 1.10]
Diag (reference: DLBCL)		
OthB	0.48 [0.25; 0.89]	0.63 [0.14; 2.91]
T	1.98 [1.23; 3.21]	3.39 [1.10; 10.5]
Other	0.80 [0.39; 1.64]	n/a
SUVbase	0.97 [0.95; 0.99]	0.99 [0.95; 1.03]
SUVint	1.08 [1.05; 1.11]	1.09 [1.01; 1.18]
AnnArb (reference: stage I)		
II	1.02 [0.48; 2.15]	0.66 [0.23; 1.93]
III	2.56 [1.32; 4.97]	0.43 [0.12; 1.56]
IV	2.44 [1.27; 4.69]	0.57 [0.19; 1.72]
LDH	2.09 [1.42; 3.08]	0.90 [0.36; 2.22]

List of abbreviations

AUC	Area under the (receiver operating characteristic) curve
BS	Brier score
BSA	Body surface area
CD20	Cluster of differentiation molecule 20
CHOP	Cyclophosphamide, doxorubicin hydrochloride, vinicristine, prednisolone
CI	Confidence interval
CONSORT	Consolidated Standards Of Reporting Trials
CoxReg	Cox regression
CP	Concordance probability
CT	Computed tomography
DLBCL	Diffuse large B-cell lymphoma
ECOG	Eastern Cooperative Oncology Group
ED	Euclidean distance
EudraCT	European Union Drug Regulating Authorities Clinical Trials
ExNod	Extranodal manifestation
^{18}F-FDG	Fluorodeoxyglucose
FL	Follicular lymphoma
FP	Fractional polynomial
FP-time	Fractional polynomial time [procedure]
HR	Hazard ratio
IBS	Integrated Brier score
iPET	Interim positron emission tomography and computed tomography
IPI	International Prognostic Index
J	Youden's J index
LDH	Lactate dehydrogenase
LogReg	Logistic regression
MAR	Missing at random
MCAR	Missing completely at random
MFP	Multivariable fractional polynomial [procedure]
MFPT	Multivariable fractional polynomial time [approach]
MNAR	Missing not at random
NHL	Non-Hodgkin lymphomas

OS	Overall survival
OthB	Other B-cell lymphomas
Other	Other lymphomas
PET	Positron emission tomography
PET/CT	Positron emission tomography and computed tomography
PETAL	Positron Emission Tomography–Guided Therapy of Aggressive Non-Hodgkin Lymphomas [trial]
PMBCL	Primary mediastinal large B-cell lymphoma
Q1 & Q3	First and third quartile
R	Rituximab
R-CHOP	Rituximab, cyclophosphamide, doxorubicin hydrochloride, vinicristine, prednisolone
RECIST	Response evaluation criteria in solid tumors
ROC	Receiver operating characteristic
SE	Sensitivity
SP	Specificity
STRATOS	STRengthening Analytical Thinking in Observational Studies [initiative]
SUV	Standardised uptake value
T	T-cell lymphomas
ULN	Upper limit of normal range
WHO	World Health Organization

List of figures

List of tables

Acknowledgements

Ich danke allen aktuellen und ehemaligen Mitarbeitern des IMIBE und des ZKSE, die mich während meiner Promotion inhaltlich oder mental unterstützt haben. Obwohl deren Anzahl zu groß ist, um sie hier in angemessener Form einzeln zu berücksichtigen, muss ein besonderer Dank den an der PETAL-Studie beteiligten Kollegen und den Mitgliedern der Arbeitsgruppe Biometrie und Bioinformatik gelten. Hervorheben möchte ich selbstverständlich meinen Doktorvater Prof. Dr. Karl-Heinz Jöckel, dessen gehaltvolle Anmerkungen in unseren unzähligen Diskussionen nicht nur statistisch-methodischer Natur ausgiebig zur Qualität dieser Arbeit beigetragen haben. Ebenso gebührt mein Dank Frau Prof. Dr. Claudia Ose, ohne die diese Dissertation sicherlich nicht zustande gekommen wäre. Darüber hinaus möchte ich mich bei Banu Demirci für die ein oder andere Einzelfallbetreuung auf administrativ-bürokratischer Ebene bedanken sowie bei Carina Emmel für einen ganz speziellen Einblick in das Thema Spiritualität. Ebenfalls danke ich Prof. Dr. Ulrich Dührsen aus der Klinik für Hämatologie, der mir in medizinischen Fragen immer ein bereitwilliger Ratgeber war und mit dem ich während der PETAL-Studie äußerst erfolgreich und in mindestens fünfundneunzig Prozent der Fälle Hand in Hand zusammenarbeiten durfte. In der Klinik für Hämatologie bin ich auch Frau Dr. Christine Schmitz zu Dank verpflichtet, die sich nicht zu schade war, mir als Hämatologin auch nuklearmedizinische Inhalte zu vermitteln und besonders in der Endphase einen sichtbar positiven Einfluss auf meine Motivation hatte. Vor allem danke ich aber meiner Mutter, meinem Vater und meinem Bruder, die mir in dieser entbehrlichen aber aufregenden Zeit stets als sicherer Hafen und unermüdlicher moralischer Fels in der Brandung zur Verfügung standen.

Curriculum vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.