STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:3401–3415

Validation, calibration, revision and combination of prognostic survival models‡

Hans C. van Houwelingen*,†

Department of Medical Statistics, Leiden University Medical Center, P.O. Box 9604, 2300 RC Leiden, The Netherlands

SUMMARY

The problem of assessing the validity and value of prognostic survival models presented in the literature for a particular population for which some data has been collected is discussed. Methods are sketched to perform validation through 'calibration', that is by embedding the literature model in a larger calibration model. This general approach is exemplified for x-year survival probabilities, Cox regression and general non-proportional hazards models. Some comments are made on basic structural changes to the model, described as 'revision'. Finally, general methods are discussed to combine models from different sources. The methods are illustrated with a model for non-Hodgkin's lymphoma validated on a Dutch data set. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Many clinical studies focus on the predictive value of some prognostic factor. A MEDLINE search for papers with prognostic or predictive in the title yields more than 20000 papers. Of course, not all of these papers deal with the actual development of prognostic models, but many do. Prognostic studies are very popular in oncology, and also occur in other chronic diseases such as cardiovascular and specialities such as transplantation. These studies can differ substantially in the goal of the research and the results reported. Some focus on finding new factors that are of 'independent' prognostic importance without specifying the actual prognostic model. Others use prognostic factors to define subgroups of patients with good or bad prognosis. A limited number come up with a prognostic model for individual patients.

Once a prognostic model for a specific disease (with a specific treatment etc.) is published, issues about its validity are raised. One particular model can hardly be expected to be generally valid for all patients world-wide. Validity will depend on the subpopulation to which the model is applied. I will study, by means of an elaborated example, the validation of a model for a well-defined

*Correspondence to: Hans C. van Houwelingen, Department of Medical Statistics, Leiden University Medical Center, P.O. Box. 9604, 2300 RC Leiden, The Netherlands.
†E-mail: j.c.van_houwelingen@medstat.medfac.leidenuniv.nl
‡Presented at the International Society for Clinical Biostatistics Nineteenth International Meeting, Dundee, Scotland, August 1998.

subpopulation of patients for whom enough data are available to carry out such a validation study. Essentially I want to answer the clinical question 'what is the value of a published prognostic study for my patients'. I restrict attention to prognostic models for survival data, but many issues easily carry over to ordinary regression models for Normal outcomes or logistic regression models for binary outcomes.

In Section 2 I discuss what is meant by a prognostic survival model. Section 3 describes the example used throughout the paper. Section 4 deals with validation and calibration, that is, adjustment of the published model by some simple rescaling. In Section 5 I discuss model revision, that is, changing and extending the literature model. In Section 6 I will deal with comparing and combining models from different sources. Finally, in Section 7 some conclusions are drawn.

2. WHAT IS A PROGNOSTIC SURVIVAL MODEL?

Strangely enough, the concept of a 'prognostic model' is far from well defined. Some authors use the term to indicate the selection of prognostic factors included in the model. Many studies are satisfied to report that 'factor X' is of independent prognostic value, implying that it can be added to an already long list of prognostic factors. That is not what I have in mind here. A general, and therefore vague, definition could be:

> *A 'prognostic survival model' is a quantification of the survival prognosis of patients based on information at start of follow-up* ($t = 0$).

Start of follow-up can be either the moment of diagnosis or the start of treatment. In clinical trials it is usually the moment of randomization, but here we are not restricted to clinical trials.

Note that I consider only fixed prognostic factors that are measured at $t = 0$. Extensions are possible to dynamic models with more stages and/or time-dependent covariates. Checking the validity of such models is more complicated than for simple fixed factor models. I will not discuss it in this paper.

From a statistical perspective the 'ideal' model is a completely specified survival model using all available information, that enables the computation of $S(t|X) = P(T > t|X)$ for all t and all possible covariate patterns X. Examples of such models are the:

(i) Cox model with regression coefficients *and* baseline survival function

$$S(t|X) = S_0(t)^{\exp(X\beta)}$$

(ii) Accelerated failure time (AFT) models

$$S(t|X) = S_0(t/\exp(X\beta))$$

(It should be stressed here that the baseline survival function is an essential part of the Cox model.)

From a clinical perspective, the ideal model is 'something simple'. Maximal simplicity is achieved if the model divides patients into 'good patients' and 'bad patients' without further specification of the survival chances of the two groups.

Models presented in the literature are a compromise between the extremes of the 'statistical ideal' and the 'clinical ideal'. Covariate information is either used by defining a prognostic index

Table I. Non-Hodgkin's lymphoma: 2- and 5-year survival probabilities.

IPI	2-year survival			5-year survival		
	Original model	Dutch Data (SE)	Calibrated model	Original model	Dutch data (SE)	Calibrated model
1	0.84	0.78 (0.03)	0.78	0.73	0.61 (0.04)	0.58
2	0.66	0.54 (0.05)	0.55	0.51	0.35 (0.05)	0.31
3	0.54	0.39 (0.05)	0.41	0.43	0.15 (0.04)	0.23
4	0.34	0.24 (0.05)	0.21	0.26	0.10 (0.03)	0.09

IPI: international prognostic index

as a weighted mean of the prognostic variables or by defining some rules for the grouping of patients. Information on survival in the model can be either completely specified or restricted to specific time-points (for example, 2-year survival probabilities). Sometimes only a ranking of subgroups is reported, occasionally accompanied by the 'relative risks' of the subgroups. (I use the term relative risk in rather a loose way. Its precise meaning depends on the type of model considered: relative hazard in Cox regression models and odds ratio in logistic regression.)

Generally speaking, validation of well-specified (quantitative) models is better defined than validation of vague (qualitative) models that only produce subgroups. Therefore, I concentrate on quantitative models in the rest of this paper, the purpose of which is both to develop statistical tools to assess the validity and value of a model published in the literature for a clinician's own patient population, and to adjust a published model to make it fit the clinician's patient population. An essential requirement for such an exercise is that the new data contain all covariates in the literature model. That presumes some consensus on the relevant prognostic factors.

3. NON-HODGKIN'S LYMPHOMA

As an example throughout this paper I take the model for overall survival for aggressive non-Hodgkin's lymphoma from Shipp *et al.* [1]. In that paper an international prognostic index (IPI) is defined based on age, Karnovsky score, Ann Arbor stage, extra nodal sites and LDH scores. The authors perform a Cox regression analysis, but finally come up with an IPI that is motivated by, but not directly derived from, the Cox regression analysis. The five risk factors are dichotomized and the IPI score just counts the number of unfavourable risk factors, thus running from 0 to 5. In a second stage the more extreme categories 0 and 1, and 4 and 5 are pooled. (The reasons for this pooling are not very clear in the paper. An impression is given that there is little difference in survival between the groups that were pooled, but there is no further information. Generally speaking, such pooling could lead to loss of information.) The resulting four categories are denoted by low, low intermediate, high intermediate and high. I will simply denote the subgroups by 'IPI = 1' to 'IPI = 4'. The information on survival given in the paper consists of Kaplan–Meier curves in the four subgroups and a table of 2- and 5-year survival probabilities. In this paper, the latter is the only information used; it is given in Table I under the heading 'Original model'.

Regrettably, the authors did not attempt to model survival in the four subgroups in any way. To demonstrate the way a full survival model can be validated and calibrated, I fitted a Weibull proportional hazards model, using only the 2- and 5-year survival probabilities of Table I. The

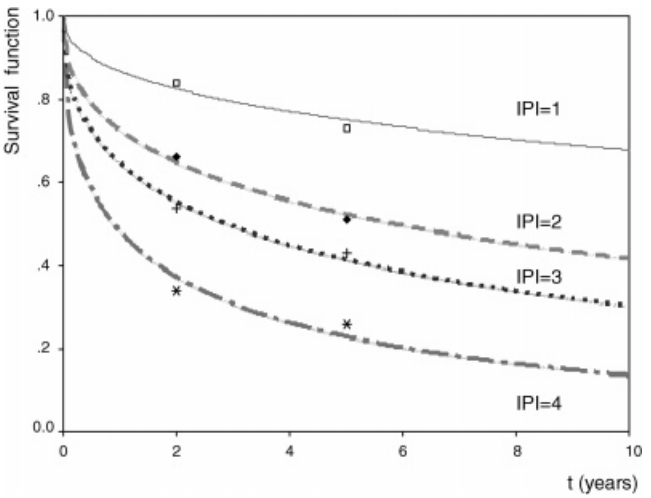

Figure 1. IPI-Weibull model The symbols indicate the 2- and 5-year survival probabilities of Table I (original model) in the four IPI subgroups. The curves give the fitted Weibull model.

general Cox model $S(t|X) = S_0(t)^{\exp(X\beta)}$ can be rewritten as $\ln(-\ln(S(t|X)) = \ln(-\ln(S_0(t)) + \mathrm{PI}(X)$, with $\mathrm{PI}(X) = X\beta$ the prognostic index of the model. The Weibull model specifies that $\ln(-\ln(S_0(t)) = \beta_0 + \beta_1 \ln(t)$. Taking the grouping variable IPI as a categorical variable we have to fit the simple model $\ln(-\ln(S_i(t_j))) = \beta_0 + \beta_1 t_j + \alpha_i$ for $i = 1,2,3,4$ and $j = 1,2$. Using simple linear regression we obtain the model

$$\ln(-\ln(S(t|\mathrm{IPI})) = -0.319 + 0.439\ln(t) + \mathrm{PI}(\mathrm{IPI}) \tag{1}$$

with $\mathrm{PI} = -1.638, -0.824, -0.514$ and 0 for $\mathrm{IPI} = 1,2,3$ and 4, respectively.

The model is denoted as the IPI-Weibull model and the corresponding graphs are shown in Figure 1. The Weibull fit may be a crude approximation but it serves well to explain the ideas of calibration and validation, although it would have been nice if the paper itself contained some model for survival in the IPI groups. The model could have been reported as 'relative risks' of 0.194, 0.439 and 0.598 for $\mathrm{IPI} = 1,2$ and 3, respectively, with respect to the baseline $\mathrm{IPI} = 4$ and a baseline cumulative hazard $H_0(t) = \exp(-0.319 + 0.439\ln(t)) = 0.727t^{0.439}$ and baseline survival function $S_0(t) = \exp(-H_0(t))$.

I want to check the validity of this model for a group of 426 similar Dutch patients (148 with $\mathrm{IPI} = 1$, 110 with $\mathrm{IPI} = 2$, 85 with $\mathrm{IPI} = 3$ and 83 with $\mathrm{IPI} = 4$), first reported in a paper by Hermans *et al.* [2]. That paper can be seen as a first validation of the IPI model in a more qualitative way. Using the IPI definitions, they obtained the Kaplan–Meier estimators in the four subgroups as shown in Figure 2.

From that Figure, Hermans *et al.* concluded that the IPI model was valid. They also considered patients that did not quite meet the IPI criteria and showed that the IPI index has some value there as well. In this paper I shall try to give a more quantitative check on the validity of the model.

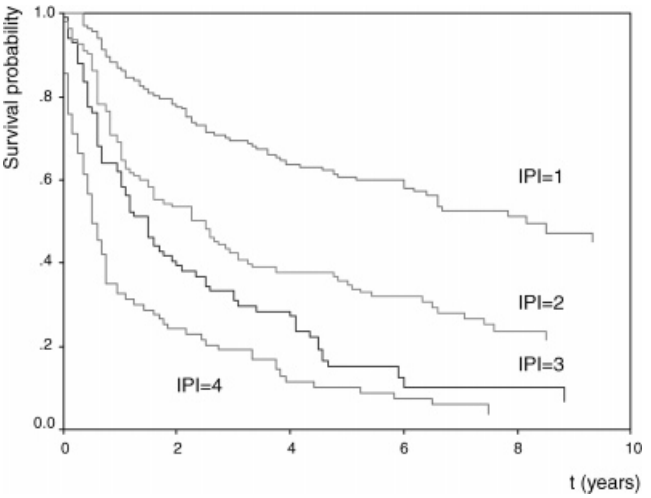

Figure 2. Kaplan–Meier curves in Dutch data.

4. VALIDATION AND CALIBRATION

The statistical problem I discuss is to check whether a given model is consistent with a set of data. Like the concept of 'prognostic model' the concept of 'validation' is not well defined. There are no generally accepted rules how validation should be carried out. The naive approach is to start from scratch, build your own model and check whether the models look the same as the given model, that is, whether it has the same prognostic factors and similar regression coefficients (and similar baseline survival function). The result is usually very disappointing at first sight, but at second sight the models can be quite similar in the sense of giving similar survival probabilities and/or patient rankings. The strategy chosen in this paper is to compare a given model with the new *data*, not with the new *model*. In very loose terms I propose to validate models by checking if the new observed data are consistent with the expected data from the model. I call this *validation by calibration*. The idea is easy to describe for simple linear regression. The literature model specifies an 'expected' value as a linear combination $X'\beta_{\text{model}}$ of the explanatory variables and the constant term contained in the covariate vector X. The new data set contains observed values Y and the same set of covariates X for each observation. The validation/calibration procedure goes as follows:

1. Plot 'observed' Y versus 'expected' $\hat{Y} = X'\beta_{\text{model}}$ for the new data set.
2. If this is scattered around the 45°-line through the origin, the model is valid without adjustment.
3. If it is a straight line with different slope and/or intercept, the model can be fixed by calibration: fit a model $Y = \alpha + \beta\hat{Y} + e$. Use $\hat{Y}_{\text{cal}} = \hat{\alpha} + \hat{\beta}\hat{Y}$ as the calibrated model.
4. If there is no correlation, it is hopeless.

In more formal language I could say that we embed the published model in a so-called calibration model that allows both the testing of the validity of the model and adjusting the model by calibration. Calibration in connection with cross-validation is used in Van Houwelingen and le Cessie [3]

and Verweij and Van Houwelingen [4] to estimate the shrinkage correction factor for overfitting. See also Copas [5] and Harrell *et al.* [6] where the term calibration is used in a similar spirit.

In this calibration set-up the model 'predictions' $\hat{Y} = X'\beta_{\text{model}}$ are considered error-free. That means that the error term e in the model above is due to both sampling variation and model misspecification. The purpose of the whole exercise is to come up with a reasonable model by using the experience from others about the relative contribution of the covariates and by estimating only two parameters, if necessary. In this way we hope to control the overfitting that is present in many models based on small data sets and many covariates.

In the setting of a survival model the observed data consist of the survival time T_i, the event indicator d_i ($d_i = 1$ if an event has occurred at T_i and $d_i = 0$ in case of censoring) and the covariate vector X_i. It is less clear how to perform a validation by calibration. What can be done depends on the way the model is reported and what kind of model is used. I discuss three different situations:

(i) calibrating estimated x-years survival probabilities in predefined subgroups;
(ii) calibrating a Cox proportional hazards model;
(iii) calibrating a non-proportional hazards model.

4.1. Calibrating estimated x-year survival probabilities in predefined subgroups

As in my example, a model is often reported as a table of x-year survival probabilities for fixed x ($x = 2$ or 5 in the example) in G subgroups defined by the covariates. We denote them as $S_{\text{model}}(x|g)$ for $g = 1, \ldots, G$. If for each patient in the new data set the follow up is at least x years, that is, if there is no censoring before x years, the model can be calibrated by logistic regression with outcome $Y = 1$ if $T > x$. If there is censoring before x years, the only thing that can be done is to estimate the x-year survival probabilities in the subgroups by the Kaplan–Meier procedure and compare the estimated $\hat{S}(x|g)$ with the 'predicted' probabilities $S_{\text{model}}(x|g)$ from the literature model. Using the $\ln(-\ln(.))$-link we get the calibration model

$$\ln(-\ln(\hat{S}(x|g)) = \alpha + \beta \ln(-\ln(S_{\text{model}}(x|g)) + e \tag{2}$$

Again the error term e comprises both sampling fluctuations in the new data set and imperfect modelling. The literature model is valid in absolute sense if $\alpha = 0$ and $\beta = 1$. If $\beta = 1$ but $\alpha \neq 0$, the literature model correctly specifies the 'relative risks' but has the wrong baseline. If both $\alpha \neq 0$ and $\beta \neq 1$ we need to estimate two parameters and the whole calibration exercise only makes sense if the number of subgroups G exceeds two. The calibrated model

$$S_{\text{cal}}(x|g) = \exp(-\exp(\hat{\alpha} + \hat{\beta} \ln(-\ln(S_{\text{model}}(x|g)))))$$

combines the new data and the literature experience and might be used to correct some 'outliers' in this data set.

4.2. Calibrating a Cox proportional hazards model

A Cox proportional hazards model consists of two parts: the prognostic index $\text{PI} = X\beta_{\text{model}}$ and the baseline survival function $S_0(t)$ or the baseline cumulative hazard $H_0(t) = -\ln(S_0(t))$. Unfortunately, the baseline information is seldom reported explicitly, but it can often be reconstructed from graphical information such as predicted survival curves for certain covariate patterns.

If we only want to check if the relative risks $RR_{a|b} = \exp(PI_a - PI_b)$ are correctly specified by the model, we can simply perform a Cox regression in the new data set with PI as single covariate, that is fitting the proportional hazards model

$$h(t|\mathrm{PI}) = h_0(t)\exp(\beta \mathrm{PI}) \tag{3}$$

with free parameters β and $h_0(t)$. If the regression coefficient $\beta = 1$, the relative risk model is valid; if $\beta \neq 1$ there is a need for calibration. This kind of calibration is very easy to carry out since it only requires the computation of the prognostic index PI for all individuals in the new data set.

If we also want to assess the validity of the whole model, it is necessary to check the correctness of the baseline survival function as well. To do so, we follow the methodology of Van Houwelingen and Thorogood [7]. If we define a proper time scale $T^* = H_0(T) = -\ln(S_0(T))$ we can represent the Cox model $S(t|\mathrm{PI}) = S_0(t)^{\exp(\beta \mathrm{PI})}$ as an accelerated failure time (AFT) model [8]

$$\ln(T^*) = (-\mathrm{PI}) + e$$

with e distributed as the logarithm of a negative exponential, that is

$$P(e < z) = 1 - \exp(-\exp(z))$$

The natural calibration model is a Weibull model on this transformed time-scale that can be written as an AFT model:

$$\ln(T^*) = \alpha + \beta \mathrm{PI} + \gamma e \tag{4}$$

The whole model is strictly valid if $\alpha = 0$, $\beta = -1$ and $\gamma = 1$. The parameter γ controls the validity of the shape of baseline. If $\gamma = 1$, the Weibull model can be replaced by the simpler exponential calibration model that retains the shape of the original baseline hazard. The parameter β controls the effect of PI and α the overall level. Notice that the Weibull model is also a proportional hazards model with regression coefficient $\beta^* = -\beta/\gamma$.

To carry out this calibration we need the baseline cumulative hazard from the literature model to compute the time transform $T^* = H_0(T)$. Some smoothing of $H_0(t)$ may be advisable. Furthermore, we need software to fit the parametric Weibull model.

4.3. Calibrating non-proportional hazards models

So far I have considered calibrating a Cox model through its representation as an AFT model. In more complicated models no simple description is available in terms of prognostic index plus baseline hazard. That is the case for stratified Cox models that have separate baselines for each stratum, non-proportional hazards models in which the effects of covariates can change over time, or AFT models with shape term depending on covariates ($\ln(T) = X\beta + \gamma_X e$).

No matter how a model has been derived, it will always specify the cumulative hazard $H_{\mathrm{model},i}(t)$ for each individual depending on their covariate information X_i. I will discuss validation/calibration using only the model cumulative hazard function. A very simple approach is to consider the calibration model

$$H_i(t) = \alpha H_{\mathrm{model},i}(t) \tag{5}$$

This is similar to fixing $\beta = -1$ and $\gamma = 1$ in the Weibull calibration discussed before and only adjusting the α-parameter. If d_i denotes the event (death) indicator and T_i the observed survival or censoring time, the straightforward estimate of α is given by

$$\hat{\alpha} = \sum d_i / \sum H_{\text{model},i}(T_i) \quad \text{with SE } (\ln(\hat{\alpha})) = 1/\sqrt{(\sum d_i)}$$

Testing for $\alpha = 1$ is just the overall test whether the model correctly specifies the total number of deaths.

A refinement is

$$H_i(t) = \alpha H^{\beta}_{\text{model},i}(t) \tag{6}$$

This can be fitted again by Weibull regression with $H_{\text{model},i}(T_i)$ as 'time' and no further covariates.

This is a calibration with only one (or two) parameter(s) for adjusting the overall survival rate (and the general shape). It relies very heavily on the specification of the hazards in the literature model. It leaves little room to include more features of the new data set in the final calibrated model. A more flexible approach would be to shrink the literature model towards 'the overall mean' in the new data. As 'overall mean hazard' we can take a smoothed version of the non-parametric estimate of the overall cumulative hazard in the new data or we can take a simple Weibull model. Let us denote this overall mean hazard by $\hat{H}_{\text{overall}}(t)$. A general calibration model is given by

$$\ln(H_i(t)) = (\alpha) + (\beta)\ln(\hat{H}_{\text{overall}}(t)) + \gamma(\ln(H_{\text{model},i}(t) - \overline{\ln(H_{\text{model},i}(t))} \tag{7}$$

(If the model were linear, $\alpha = 0$ and $\beta = 1$ would do. There is some need for estimating α and β as well due to the non-linearity of the model.)

For a proportional hazards model this reduces to a Weibull regression with a time transform obtained from the actual data instead of from the model. I will not elaborate this further, but it gives an idea of how calibration can be achieved for general models.

4.4. Non-Hodgkin's lymphoma continued

The 2- and 5-year survival probabilities derived from the Kaplan–Meier estimates in the Dutch data set are given in Table I under the heading 'Dutch data'. We apply calibration model (2). 'Observed' and 'expected' are shown on the $\ln(-\ln(.))$-scale in Figure 3.

We obtain good calibration with $\beta = 1$ for both $x = 2$ and $x = 5$, $\alpha = 0.37$ for $x = 2$ and $\alpha = 0.56$ for $x = 5$. The conclusion is that the IPI model has correct 'relative risks' but wrong 'baseline'. Note that no attempt is made at this stage to link the results for different values of x. The survival probabilities derived from the calibration model are given in Table I under the heading 'Calibrated model'. They differ substantially from the original IPI survival probabilities, but are very similar to the original 'Dutch data' estimates; this is because the number of groups G is quite small ($G = 4$) and the relative risks in the Dutch data are surprisingly similar to the ones in the IPI model. Apparently, the only effect of the whole calibration exercise is the correction of the outlying observation at $x = 5$ in the group IPI = 3.

To get a first impression of the performance of the IPI system for the Dutch survival data I perform a Cox regression with IPI as categorical covariate. The results are given in Table II. These coefficients are very similar to those in the Weibull-IPI model (1). Actually, there is no need to compute these estimates since we can directly feed the PI score into the Cox regression of model (3). That gives a regression coefficient of $\hat{\beta}_{\text{PI}} = 1.034$ with (SE 0.102). It looks nearly perfect. The

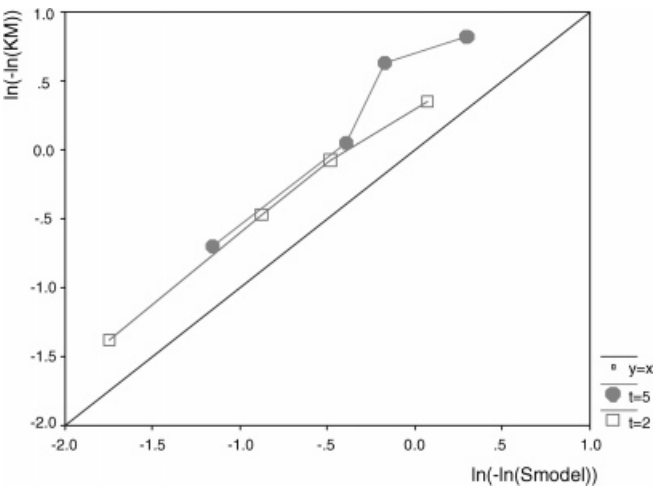

Figure 3. Calibration on $\ln(-\ln(.))$-scale.

Table II. Non-Hodgkin's Lymphoma Cox regression on the international prognostic index (IPI).

Category	$\hat{\beta}$	SE
[IPI = 1]	−1.67	0.17
[IPI = 2]	−0.92	0.16
[IPI = 3]	−0.48	0.16
[IPI = 4]	0	

Table III. Non-Hodgkin's lymphoma Weibull calibration.

Parameter	Estimate	SE
α	−0.24	0.06
β	−0.68	0.07
γ	0.65	0.03

simple IPI-Weibull model gives the correct relative risks for the Dutch data. The null hypothesis $\beta_{\mathrm{PI}} = 1$ need not be rejected but how about the baseline? Figure 4 shows that the IPI baseline and the baseline in the Dutch data are definitely different.

To check the whole model we perform the Weibull calibration of model (4). The time transform implied by the IPI-Weibull model is $T^* = H_0(T) = 0.727T^{0.349}$.

The result of the Weibull regression with time variable T^* and PI as single prognostic index is given in Table III. Since γ is clearly smaller than one, it can be concluded that an exponential model (with $\gamma = 1$) does not fit well. The baseline hazard cannot simply be adjusted by multiplying by

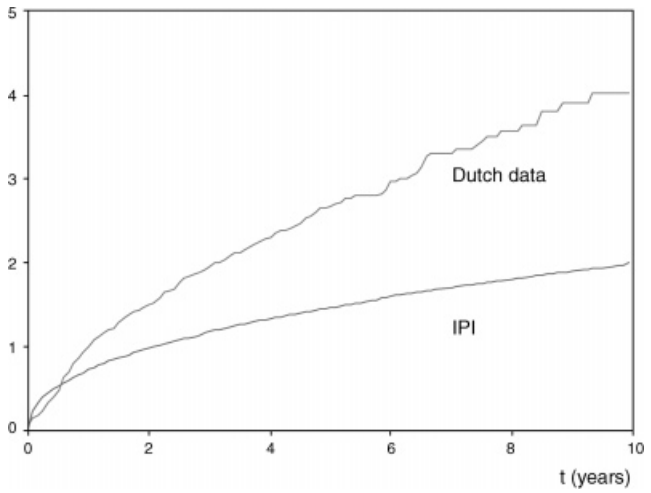

Figure 4. Baseline cumulative hazards.

a suitable constant. In the proportional hazards interpretation of the Weibull model the regression coefficient for the prognostic index is $-\hat{\beta}/\hat{\gamma} = 0.677/0.649 = 1.043$. Thus, once again, we see that the effect of PI is correctly quantified in the IPI model.

The lesson that can be learned from this exercise is that a satisfactory model for the Dutch data can be obtained by estimating only *three* parameters, two of which are needed to obtain the baseline survival and one to calibrate the IPI effect.

The IPI-Weibull model is a proportional hazards model, but just for the sake of the example I show some results of the application of the methods of Section 4.3. If we perform the one-parameter calibration of model (3) for the Dutch data and the IPI-Weibull model we obtain $\hat{\alpha} = 301/194.6 = 1.547$, $\ln(\hat{\alpha}) = 0.436$ $(\mathrm{SE}(\ln(\hat{\alpha})) = 0.058)$. Obviously, $\alpha > 1$, indicating that the overall survival is worse in the Dutch data. The calibrated survival curves, together with the Kaplan–Meier estimators, are shown in Figure 5.

The calibrated survival curves more or less agree with the Kaplan–Meier estimators, but there are very clear 'crossings' indicating room for further calibration. Fitting the two-parameter model (6) gives $\hat{\alpha} = 1.9$ and $\hat{\beta} = 1.4$. The model fits significantly better, but the graphs (not shown) still show a considerable lack of fit.

5. MODEL REVISION

Returning to the Cox model, there might be a need to revise a model in the sense that either the weights of the covariates in the regression equation are corrected or that new covariates are added. This comes close to building a new model on the new data. I advocate being conservative here as well and taking the literature model as starting point.

To be more specific, consider a literature-based Cox regression model with k covariates and prognostic index $\mathrm{PI} = \sum_{i=1}^{k} \beta_{\mathrm{model},i} X_i$. The calibration model (3) assumes that the relative influence of each covariate is correctly specified. However, there might be reasons why the effect of some of the covariates might have changed, for instance a slight change in definition or a new measuring

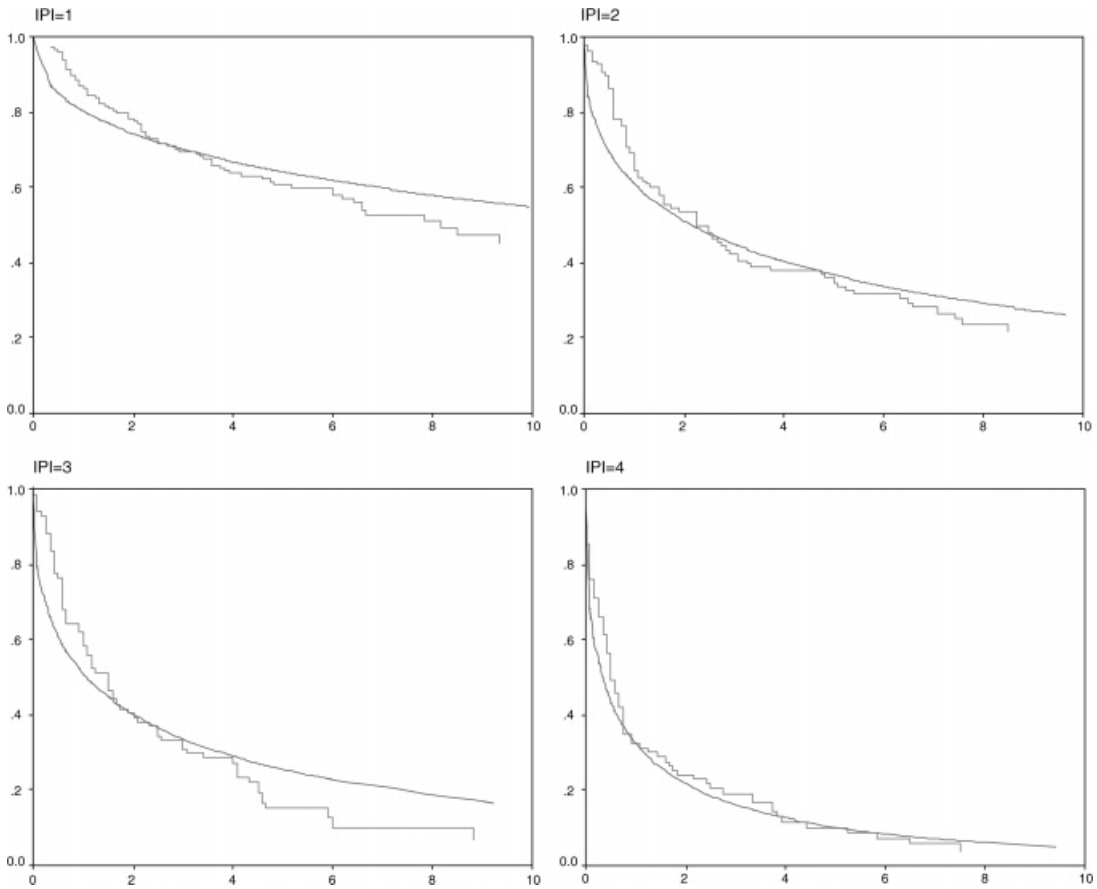

Figure 5. Comparison of Kaplan–Meier estimators with simple calibration of the IPI-Weibull model.

device. Also, if the covariate distribution in the new data set is markedly different from the one in the literature sample, all kind of mechanisms may lead to different regression parameters in the new data set. Simply re-estimating the regression coefficient might replace a reliable but slightly biased estimate by an unbiased but very unreliable one. I suggest to start with the calibration model (3) and only change the coefficient of some explanatory variable X_i if necessary. That can be achieved by a stepwise forward procedure in which the prognostic index PI is included first and the components X_i are only added to that model if significant. Thus, the model used in this revision process is

$$h(t|X_1,\ldots,X_k) = h_0(t)\exp\left(\alpha\sum_{i=1}^{k}\beta_{\text{model},i}X_i + \sum_{i=1}^{k}\beta_{\text{corr},i}X_i\right) \tag{8}$$

Here, the parameters $\beta_{\text{model},i}$ are given by the literature model and fixed, while the parameters α and $\beta_{\text{corr},i}$ and the baseline hazard $h_0(t)$ have to be estimated from the data. The full model is not identifiable and is equivalent to building a completely new model. Fitting the full model does not make sense, but the model is very well suited for a forward model construction that

starts with only the α-parameter and corrects the regression coefficients of the covariates only if necessary. This model re-estimates the baseline hazard. As argued before, it is also possible to use the baseline hazard from the literature in the Weibull calibration model (4). It is not hard to extend that model in a similar way.

If new explanatory variables, $X_{k+1}, X_{k+2}, \ldots$ become available that have not been included in the literature model, a similar procedure can be followed. Again the literature model is the starting point and new covariates are only added if they significantly improve the calibrated literature model. The modification of model (8) is straightforward

$$h(t|X_1,\ldots,X_k,X_{k+1},\ldots) = h_0(t)\exp\left(\alpha\sum_{i=1}^{k}\beta_{\text{model},i}X_i + \sum_{i=k+1}\beta_i X_i\right) \tag{9}$$

In this way the maximal parsimony is achieved in obtaining a model for the new data and overfitting, the curse of all statistical modelling, is hopefully prevented.

Another issue, half-way between calibration and revision, is checking and correcting the proportional hazards assumption in Cox models. The proportionality can be checked by studying the interaction between the index PI and a suitable function $f(t)$ as time-dependent covariate in the model

$$h(t|\text{PI}) = h_0(t)\exp(\beta\text{PI} + \gamma\text{PI}f(t)) \tag{10}$$

Convenient choices for $f(t)$ are $f(t)=t$ or $f(t)=\ln(t)$. Another possibility is to let the γ-parameter in (4) depend on the prognostic index PI. Checking for non-proportionality within Cox models should be a part of the model building process. If models are presented as proportional hazards models and apparently fail the proportionality test, the non-proportionality should be included in the model. Building a non-PH model out of the given model goes beyond simple calibration; it is true model building. The validity of the new non-PH model has to be checked in a new validation round.

5.1. Non-Hodgkin's lymphoma continued

In the IPI example we can check whether the IPI score can be improved by reweighting the components in the IPI index. I consider a Cox model with the PI index based on IPI as the only covariate as the null model and test whether inclusion of any of the dichotomous building blocks of IPI leads to significant improvements. It appears that only the dichotomy [AGE > 60] is statistically significant with positive coefficient and the other ones are not. Apparently, the role of age is underestimated in the IPI system. This was already hinted at in the original paper. Besides the IPI index a stratified index was also presented; this means one index for patients under 60 years and another for patients over 60 years.

Checking the proportional hazards assumption, I found no significant interaction with time itself, but a highly significant interaction with ln(time). This shows again that checking for non-proportionality is not straightforward. A suggestion of non-proportionality is already seen in Figures 1 and 2.

6. COMPARING AND COMBINING COMPETING MODELS

Different literature sources may give different models that can be applied to the new data. Two questions can be raised: 'which model is best?' and 'how can the models be combined?'.

Consider the situation with two models specifying survival functions $S_{1i}(t)$ and $S_{2i}(t)$, respectively, for each individual in the new data set with corresponding hazard functions $h_{1i}(t)$ and $h_{2i}(t)$ and cumulative hazard functions $H_{1i}(t)$ and $H_{2i}(t)$. For the present I ignore the calibration problem and discuss how we can compare the models and combine them into one model. Comparison can be easily done by comparing the log-likelihoods

$$ll_m = \sum_i d_i \ln(h_{mi}(T_i)) - H_{mi}(T_i)) \quad \text{for } m = 1, 2$$

The model with the largest log-likelihood is the winner.

If the models are calibrated before being put to the test, an Akaike correction

$$ll = ll\text{-number of calibration parameters}$$

can be used before comparing the models.

Notice that we run into trouble here if we use the usual Cox models with jumping hazards because the new data will occur at 'new' time points at which the cumulative baseline hazard is constant, and, hence, the estimated hazard $h_{mi}(t) = 0$ and $ll_m = -\infty$. Therefore, the hazards have to be smoothed. Smoothing of the hazard is a neglected part of the model building process and the success of a model may heavily depend on the way the baseline hazard is smoothed.

There are several ways of combining models. I discuss three possibilities:

$$\text{mixture model} \qquad S_i(t) = (1-\theta)S_{1,i}(t) + \theta S_{2,i}(t) \tag{11a}$$

$$\text{additive cumulative hazard model} \qquad H_i(t) = \alpha_0 H_{1,i}(t) + \alpha_1 H_{2,i}(t) \tag{11b}$$

$$\text{multiplicative cumulative hazard model} \qquad \ln(H_i(t)) = \beta_1 \ln(H_{1,i}(t)) + \beta_2 \ln(H_{2,i}(t)) \tag{11c}$$

The first two models are easy to fit by EM or similar algorithms. I briefly discuss how that can be done.

For the mixture model (11a), define

$$Z_i = \theta S_{2,i}(t) h_{2,i}(T_i)^{d_i} / [\theta S_{2,i}(T_i) h_{2,i}(T_i)^{d_i} + (1-\theta) S_{1,i}(T_i) h_{1,i}(T_i)^{d_i}]$$

for each individual, with θ as estimated in the previous step. The new θ is estimated by the mean of the Z's. Iteration leads to the maximum likelihood estimator of θ.

For the additive cumulative hazard model (11b), the iteration is based on computing

$$W_i = \alpha_2 h_{2,i}(T_i) / [\alpha_2 h_{2,i}(T_i) + \alpha_1 h_{1,i}(T_i)]$$

for each individual and estimating the new α's by

$$\alpha_2 = \sum W_i d_i / \sum H_2(T_i) \quad \text{and} \quad \alpha_1 = \sum (1 - W_i) d_i / \sum H_1(T_i)$$

The multiplicative model is harder to fit. With proportional hazards models

$$H_{m,i}(t) = H_{m,0}(t) \exp(\mathrm{PI}_{m,i})$$

Table IV. Non-Hodgkin's lymphoma: Comparing and combining models.

Model	Loglikelihood	Parameters	
S_1	−717.82		
S_2	−696.3		
Mixture	−686.62	0.38	0.62
Additive	−684.73	0.38	0.62
Multiplicator	−663.33	0.33	1.07

the new model (11c) reads

$$\ln(H_i(t)) = \beta_1 \ln(H_{1,}(t)) + \beta_2 \ln(H_2(t)) + \beta_1 \mathrm{PI}_{1,i} + \beta_2 \mathrm{PI}_{2,i} \tag{12}$$

A simplification in the spirit of model (3) would be to consider a simple Cox model with PI_1 and PI_2 as covariates. However, that ignores all information about the shape of the baseline hazard.

Notice that in the models (11b) and (11c) that combine the hazards, calibration is automatically built in since no restrictions are put on either $\alpha_1 + \alpha_2$ or $\beta_1 + \beta_2$.

Construction of these combination models has some similarity with the discussion about model uncertainty [9,10]. However, in our procedure one further step is made by creating new models from existing competing models.

6.1. Non-Hodgkin's lymphoma concluded

I do not have two literature models available. To demonstrate all the methods suggested above I consider two models for the Dutch data: the overall Weibull model in the new data set

$$\ln(S_1(t)) = -\mathbf{0.980} + \mathbf{0.617}\ \ln(t) \tag{13a}$$

and the IPI model with calibration according to model (5)

$$\ln(-\ln(S_2(t|I\mathrm{PI})) = -0.319 + 0.439 \ln(t) + \mathrm{PI(IPI)} + \mathbf{0.436} \text{ with PI(IPI) as in (1)} \tag{13b}$$

Both models are partly based on the new data. The former contains two parameters estimated from the new data and the latter has one such parameter (parameters estimated from the data are indicated by bold print). The results are given in Table IV.

First of all, we notice that model (13b) performs much better than the simple Weibull model that ignores all covariate information. As far as combination of models is concerned, the multiplicative model performs best. Mixture modelling does not perform very well here. Notice that all three ways of combining models do not require any specific model; they are natural extensions of the calibration methods of models (5) and (6).

7. GENERAL CONCLUSION

If a reliable model is available in the literature, it is usually quite easy to turn that into a good model for new data by estimating only a few calibration parameters without having to start from scratch and build a completely new model with all the dangers of overfitting and lack of reproducibility. Models with only a few parameters are quite stable. Following the calibration approach,

the variances of the estimated models are kept under control, may be at the cost of introducing some bias. The urge to improve the model by redefining the prognostic index or adding new prognostic factors should be repressed. The best thing statisticians can do in model building is to keep the number of new parameters as low as possible [6, 11]. Paraphrasing an American commercial, statisticians should be 'frugal fitters'. In trying to apply models from the literature to one's own data, it is always a nuisance to find out that the published models are not specific enough to produce estimated survival probabilities for all patients at all time-points. It is the duty of the biostatistician involved in reporting the prognostic model to give all the information needed to build further on their model. For Cox models that should also include the baseline hazard or survival rate, if possible smoothed somehow or given in an approximate functional form using (fractional) polynomials [12], exponentials, rational functions or something similar.

ACKNOWLEDGEMENTS

The Dutch data on non-Hodgkin's Lymphoma were made available by the Comprehensive Cancer Center West (IKW), Leiden, The Netherlands.

REFERENCES

1. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive Non-Hodgkin's Lymphoma. *New England Journal of Medicine* 1993; **329**:987–994.
2. Hermans, J, Krol ADG, van Groningen K, Kluin PhM, Kluin-Nelemans JC, Kramer MHH, Noordijk EM, Ong F, Wijermans PW. International prognostic index for aggressive Non-Hodgkin's lymphoma is valid for all malignancy grades. Blood, 1996; **86**:1460–1463.
3. Van Houwelingen JC, le Cessie S. Predictive value of statistical-models. *Statistics in Medicine*, 1990; **9**:1303–1325.
4. Verweij PJM, van Houwelingen HC. Cross-validation in survival analysis. *Statistics in Medicine*. 1993; **12**:2305–2314
5. Copas JB. Cross-validation shrinkage of regression predictors. *Journal of the Royal Statistical Society. Series B* 1987; **49**:175–183.
6. Harrell FE, Lee KL, Mark DB, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
7. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995; **14**:1999–2008.
8. Keiding N, Andersen PK, Klein JP. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* 1997; **16**:215–224.
9. Buckland ST, Burnham KP, Augustin NH, Model selection: An integral part of inference. *Biometrics* 1997; **53**:603–618.
10. Chatfield C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, *Series A* 1995; **158**:419–466.
11. Ye JM, On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* 1998; **93**:120–131.
12. Royston P, Altman DG, Regression using fractional polynomials of continuous covariates — Parsimonious parametric modeling. *Applied Statistics* 1994; **43**:429–467.

