External Validation

Terry Therneau

4 October 2023

1 Background

(Note: this chapter still has a lot places that need to be fleshed out, but the basic structure is sound.)

1.1 What does it mean to validate a model?

> If you don't know where you are going, you might end up somewhere else. Yogi Berra

Validation is a word that is often used but rarely defined, so much so that it has become essentially meaningless without further clarification. We will separate 3 meanings

- Software validation. This is discussed in appendix **??**.
- Internal validation: checks that further examine the fit of a model, using the original data set. These form an extension of the familiar trio of functional form, proportional hazards, and undue leverage which have been discussed earlier. Several of the methods found in this chapter will have been encountered earlier in the book, in that context.
- External validation. Application of a model to new data set, one which was not used in the model's construction.

External validation tries to answer the question of whether a particular model is applicable outside of the data set it where it was developed. The most critical step, before computing metrics, is to carefully think through what we mean by "applicable". In what sense does a given model provide, or not provide, a useful prediction? Any careful answer to this will always be problem specific. One consequence is that this chapter will contain ideas on how to *think* about the problem and methods, but no final checklist with a "best" solution.

This key consideration is surprisingly absent from most papers on the topic of survival model validation; two notable exceptions which have greatly influenced our thinking are Korn and Simon [**?**] and Alman and Royston [**?**]. As an example from the first, suppose that we were using $t - \hat{t}$ as the metric for goodness of prediction, the difference between the observed and predicted time to death. Should an (observed, predicted) pair of (.5y, 1y) be treated as the same size error as (10.5y, 11y)? In a study of acute disease we might consider the first pair to be a much more consequential error than the latter. For example, in a cancer in which 5 year survival is considered a "cure", one might consider any pair (a, b) with both a and b greater than 5 to be

no error at all. Another example would be a model which will be used to guide a transition to palliative care, e.g., if the predicted probability of death within 2 months were over 90% then further treatment is considered to be futile. In this case distinguishing between those with an early death rate of 10, 30 or 50% is immaterial for the planned application.

In general, the chosen validation metric needs to be *fit for purpose*. One of the most common outcomes of such a consideration, in our experience, will be to restrict the upper range of the assessment to a threshold τ; this could be 3 months or 10 years depending on the case at hand, and will be a consideration in each of our examples. A companion issue is that the validation data set might have very little information beyond some point; if only a handful of the validation subjects have more than 4 years of follow-up, for instance, then an assessment of predictive ability at 10 years will be a fool's errand, whether or not that were a target of interest.

1.2 Data checks

A few checks of the validation data should be done early in the process. These include a Kaplan-Meier plot of the outcome, overall or by subgroups, numerical and/or graphical summaries of each of the predictor variables found in the reference model, and a refit of the reference model using the new data. The goal of this is not validation per se, but to uncover simple data issues such as a lab test in different units, a different range of subjects, or a change in time scale.

2 Targets

The next question is what estimate to use as a numerical target for our efforts. For a survival outcome there are four obvious targets: the time to event t_i for each validation subject along with a predicted time $\hat{t}_i$, the observed and prediced number of events up to the target time, the observed and predicted survival at the target time, or the simple linear predictor itself $\hat{\eta}_i = X_i\hat{\beta}$. (For a Cox model $\exp(\hat{\eta})$ will be the hazard ratio). A second choice that needs to be made is whether to use a summary based on relative prediction or absolute prediction, commonly denoted as *discrimnation* and *calibration*.

Discrimination measures whether the outcomes for two subjects are in the same relative order as the predicted values for that pair of subjects. In many cases this is sufficient, for instance if the prediction will be used to stratify treatment assignment of subjects in a new trial, then only the relative ordering of their risk is required. Individual patient counseling, on the other hand, will require absolute predictions for that patient. Likewise, predicted sample size for a trial with a survival endpoint depends on the eventual *number* of events that will occur, a task that also requires absolute predictions.

A particular advantage of discrimination for a Cox model is that we don't need to decide between $\hat{t}$, $\hat{N}$, $\hat{S}$ or $\hat{\eta}$: the order of any one of these completely determines the order of all the others, i.e., if $\hat{\eta}_i > \hat{\eta}_j$ then $\hat{S}_i(t) < \hat{S}_j(t)$ for all times t. Most literature reports will omit the baseline hazard of a Cox model fit, and without this only $\hat{\eta}$ can be computed; this allows for discrimination but not calibration. This will be discussed more fully in the examples.

3 Discrimination

The primary tool for discrimination is the concordance statistic

$$C = Pr(y_i > y_j | \hat{y}_i > \hat{y}_j) \tag{1}$$

where for this paragraph we are using y and $\hat{y}$ as generic stand-ins for whichever outcome is the focus. Basic features of concordance are:

- Kendall's τ-a, Kendall's τ-b, Goodman's γ, and Somers' d (or D or D_{xy} depending on the reference) represent different choices of how to deal with tied values in y or $\hat{y}$, but target the same population quantity. These measures range from -1 to 1.
- The concordance C ranges from 0 to 1, and $C = (D_{xy} + 1)/2$, a rescaled version of Somers' d.
- If y is a 0/1 binary outcome, then $C =$ the area under the receiver operating curve (AUROC).
- Harrell [**?**] proposed an extension of C to censored outcomes. Essentially, we only score those pairs of observations where the ordering is certain. The pair of times (10+, 20) for instance is one such, censored at 10 and and event at 20. We do not know if the first observation's event time will be before or after the second. (Note that censoring does not afflick predicted values $\hat{y}$.)

A natural interpretation of concordance is to consider the evaluation of an oracle — a subject matter expert, a statistical prediction tool, a tarot deck, ...— by presenting subjects 2 at a time. For each pair count whether the oracle correctly predicts the order of their final outcome, concordance is the overall fraction correct. Pairs with a tied outcome are not presented, as they are uninformative with respect to evaluation. If the oracle cannot decide give a score of 1/2. A random die throw will have $C = 1/2$ and a perfect oracle $C = 1$, values $< .6$ from a statistical model are normally not very impressive. Values $< 1/2$ are possible (consider some political pundits),

It turns out that there is a very close association between the concordance and standard test statistics for survival data. The latter can be written as a sum of terms, one for each event or death time

$$Q = \sum_{i \in d} w(t_i)(x_i(t_i) - \overline{x}(t_i)) \tag{2}$$

where t_i is the event time of the death, w is a time-dependent weight, x_i is the covariate vector of the event (possibly time dependent) and $\overline{x}$ is the mean covariate vector over all those at risk.

- With $w =1$ Q is the score statistics for a Cox model.
- With x a 0/1 treatment indictor and $w(t) = n(t)$, the number at risk at time t, Q is the Gehan-Wilcoxon test.
- Let $x_i(t) = r_i(t)$ be the rank of the risk score η_i amongst all those at risk at time t, $0 \le r \le 1$ and $w = n(t)$. Then Q is the numerator of the concordance C.

- Many variations of the Gehan-Wilcoxon have been made over the years, including $w(t) = S(t)$ (Peto and Peto), $w(t) = S/G$ (Schemper), $w(t) = 1$ (log-rank), the $\gamma - \rho$ family (Fleming and Harrington), the Tarone-Ware, and others. We can apply any of these weights to the concordance as well. The modified concordance suggested by Uno [**?**], in fact, is exactly the Schemper weights.

- As consequence of the Cox model equivalence, all these versions of C immediately generalize to time-dependent covariates.

This set of interconnections is not widely appreciated. Several of them became clear in programming the survival package, i.e., deja vu moments we we realized that "I have seen this computation before", and is the motivation for the quote above.

If the decision is made to limit an assessment of validation to the time interval $0 - \tau$ per the considerations above, the concordance statistic can be similarly limited. This is a direct option in some software, or can be accomplished by artificially censoring any observed validation times that are greater than τ. In our experience, large differences between the various flavors of the concordance statistic only arise when the number at risk becomes small; differences are often negligible when an upper limit has been applied. This parallels the older debates about a "best" test statistic, where the final result has been that, outside certain edge cases, there is little practical difference.

4 Calibration

Multiple metrics have been proposed as measurements of calibration. The key concern, in all cases, is how to best handle censored observations in the validation data set. As a trivial example, consider using $cor(t_i, \hat{t}_i)$ as a measure, the correlation between the observed and predicted survival time. But what do you do with someone censored at $t_i = 1.5$ years with a prediction survival of $\hat{t}_i = 3$?

I organize the methods into 4 groups, and will consider each in turn.

1. Comparison of the total number of observed deaths in the validation group to the expected number predicted by the model, each subject's prediction is taken at the time of their last follow-up (or event). This are closely related to the standarized incidence ratio SIR, common in epidemiology, but less well known in statistics. It also has a close connection to counting processes.

2. Binomial methods, which compare the observed probability of survival to a selected time τ to the expected proportion. The methods are a 2 step process: first modify the censored data so as to obtain uncensored binomial observations, then apply standard approaches to the modified data. The first step can be based on inverse probability of censoring weights (IPCW), imputation, or psueudo values. This is by far the most common approach due to its familiarity.

3. Survival methods. A survival model is fit to the validation data set; this deals with censored values in a natural way. Predicted values from this new model compared to the predictions from the target model.

4. Ignore the censoring; a bad idea which is unfortunately not uncommon.

4.1 Counting process approach

The focus of this approach is to compare observed to expected events. The underlying approach is based on 4 simple ideas. Assume that $F(t)$ is the cumulative distribution function for the continuous random variable T. Then

1. $F(T) \sim U(0,1)$. This is a standard statistical result.
2. $\Lambda(T) \sim \exp(1)$, the cumulative hazard has an exponential distribution. This follows from the fact that the hazard function satisfies $H = -\log(1-F)$, and that -log of a uniform random variable is exponential.
3. If the censoring time C is independent of T and $Y = \min(C, T)$ is the censored survival time, then $\hat{\Lambda}(Y)$ is distributed as a censored exponential.
4. If a model is correct, then $\hat{\Lambda}_i(t_i)$, will also follow a censored exponential, where $\hat{\Lambda}$ is the predicted cumulative hazard from the statistical model.

Using the last statement above, along with the relationship between the Poisson and exponential distributions, allows for a simple computational tool for validation. That is, a simple Poisson GLM model with the per-subject event count as the response (the 0/1 status variable), and $-\log(\hat{\Lambda}(t_i))$ as the offset.

Berry [**?**] derives the above approach, applied to the case where $\hat{\Lambda}$ is based on population death rate tables, and shows that the Poisson likelihood is correct up to a constant factor. A more modern derivation can be based on counting process theory; A validation approach $\hat{\Lambda}$ from a fitted Cox model was outlined in [**?**]. The intercept term from the GLM model is then an estimate of $\log(O/E)$.

This ratio O/E, where O= observed number of events up to τ and E = expected number= $\sum \hat{\Lambda}_i(\min(t_i, \tau)$ is known as a standardized incidence ratio (SIR) and has long usage in population science. Keiding and Clayton [**?**] trace the SIR's history through multiple disciplines and journals for over 200 years. It is, however, less familiar in the survival literature.

4.2 Binomial methods

IPC weights The redistribute to the right (RTTR) algorithm was elaborated in section **??**. If we apply the RTTR algorithm to the validation data, up to time τ, the result is that all of the observations with unknown τ year outcome now have a weight of zero and can be dropped. The final data set now has an ordinary binomial outcome d_i =1 for a death and 0 for alive, with remaining weights $w_i \geq 1$. If the censoring is independent of outcome, we expect the results of this process to be unbiased.

This is equivalent to using inverse probability of censoring (IPC) weights, based on a reversed Kapan-Meier estimate of $G(t)$ the censoring distribution. More generally, per-subject estimates G_i can be obtained using a sub-model to predict the censoring probability. This will be more resistant to covariate dependent censoring.

Imputation The IPC method removes subjects censored before τ, and alternative is to replace their outcome with an imputated value. In the simplest form, for someone censored at t use the conditional probability of survival to τ of $p_i = S(\tau)/S(t_i)$ based on the usual Kaplan-Meier.

Either replace the observation's response with p_i or two pseudo-observations with weights p_i and $1 - p_i$ and outcomes of 0 and 1. Alternatively, make use of covariates to fit a survival model, and use predicted values from that model.

Pseudo-values This approach replaces the 0/1 response for all subjects with a pseudo-value, which is based on the influence of each subject on the estimated survival at τ. The response time t^* is now an uncensored continuous variable (binomial like in its distribution).

Once the 'binomial' response has been created, then standard methods for validation of binomial data can be employed. This includes summaries of sensitivity, specificity, postive and negative predicted value, and the AUROC. (This will not be the same as the concordance). Beyond these simple summaries, regression methods can be used to evaluate how the prediction model works across the range of predicted risks. Do those with lower predicted risk have higher survival, and in the pattern that is predicted? This can be approached by including the predicted probability as a term in binomial regression models, reminiscent of the Hosmer-Lemishow approach.

If the reference model satisfies proportional hazards, e.g. either a Cox or Weibull fit, then we know that

$$\log(-\log(\hat{p}_i(t))) = \eta_i + \log(\Lambda_0(t))$$

where $\hat{\eta} = X\hat{\beta}$ is the linear risk score based on the validation data set's covariates X and coefficients $\hat{\beta}$ from the reference model. This leads directly to the idea of fitting a binomial regression model to the weighted d_i values using a complimentary log-log link, and with η_i as the single predictor. A series of models that treat η as an offset term, as a linear predictor, or in a more flexible way such as a spline are often referred to as assessments of *mean*, *weak*, and *moderate* calibration. The first model can be used to get confidence intervals for the observed/expected ratio, a coefficient of 1.0 in the second shows that actual risk rises in magnitude as predicted by the model, and the nonlinear fit can reveal more subtle changes.

Another option is to compute the ordinary R^2 statistic based on the 0/1 outcome d_i and the predicted values $\hat{p}_i(\tau)$ from the reference model and $\hat{p}_0(\tau)$ a model with no covariates..

$$\begin{aligned} R^2 &= 1 - \frac{\sum w_i(d_i - \hat{p}_i)^2 / \sum w_i}{\sum w_i(d_i - \hat{p}_0)^2 / \sum w_i} \\ &= 1 - \frac{A}{B} \\ p_0 &= \frac{\sum w_i d_i}{\sum w_i} \end{aligned}$$

When the first step was based in the RTTR, then the numerator A of the fraction is also known as the Brier score, and A/B is Kattan's index of prediction. From the equivalence of the RTTR and Kaplan-Meier, we also know that p_0 above is the value of the overall KM of the validation data set, at τ.

4.3 Re-modeling appoach

A very useful approach described in Austin et al [**?**] is to combine imputation and modeling into a single step. That is, fit a survival model to the validation data, using the linear predictor

from the reference model as the covariate; as this deals with censoring in a natural way. Most often this second step is done using a Cox model, of course, as it is the most popular survival model. If this second model is a correct model for the outcome, then it also will be immune to the censoring pattern, if censoring is a predictable process.

Assuming that the cutoff time τ was chosen with serious thought, as should always be the case, then any events after τ are, strictly speaking, irrelevant. A modeling approach that uses all the events will be subject to a classic bias/variance trade-off: using information after τ can provide more precision, but only to the extent that the reference model remains true beyond that chosen cutoff. In the figure, this is particularly true for the uncensored case: about 15% of the simulated events occur after time 4, censoring systematically reduces that proportion.

5 Examples

5.1 Simulation

To better understand the impact of censoring we employ a set of simulation data sets. The reference model is based on a fit using 2982 primary breast cancer patients from the Rotterdam tumor bank. Simulated times for samples of n subjects are created from that model; data sets 1–5 share identical covariates and true survival times, and differ only in the censoring that is applied.

1. No censoring, used as a reference.
2. Random administrative censoring, $U(a, b)$.
3. Risk specific censoring using an exponential censoring time, chosen so that those at the 80th percentile of risk have approximately 2x the censoring rate of those at the 20th percentile.
4. As in 2, but reversed. Those with a lower death rate have higher censoring.
5. Informative censoring: a random 10% of the subjects are removed from the study 2–5 months before their death, along with administrative censoring that applies to all.

For each of the data sets we targeted 50% censoring by 4 years, and all our illustrations will use $\tau = 4$.

For all these cases the simulated survival times are drawn from the true model, and a perfect validation would show that they agree with the reference model perfectly. The question is whether censoring obscures this fact. There are two reference models, the first a Cox model fit and the second a log-logistic accelerated failure time (AFT) fit.

Figures 1 and 2 show results for the RTTR in the upper most lines, for each of the 5 censoring patterns. For both no censoring and random censoring the estimates for mean calibration (observed/expected) and calibration slope are close to 1, as expected. When there is risk-dependent censoring, then biases appear: either an upward bias in the mean in case 3 or downward in case 4. (I'd like to understand why this particular pattern occurs, but have not yet worked it out. I knew that resistance to non-random censoring was not guaranteed, but didn't know what to expect.) An appropriate response to this bias, not shown here, would be to use more general inverse probability of censoring (IPCW), based on a model for censoring that included the risk score as a predictor.

For the informative censoring of case 5 the number of *reported* deaths in the validation data set is systematically too small. None of the 4 methods can overcome this systematic bias, nor do we expect them to.

Figures 3 and 4 show results when the true model is a log-logistic AFT. For the RTTR there are no obvious changes in the pattern.

In real datasets it will be important to look for non-linear patterns as well. Figure 5 shows results from fitting a 3 degree of freedom natural spline to two simulation instances. In the left two panels the underlying model is true, and in the right two the true risk was subjected to an upper threshold. Two choices for the non-linear plot are to show the usual logistic results, as displayed in the top two panels. The horizontal axis is the linear predictor from the reference model, which was used as the covariate in the model, and the vertical axis is the predicted value from the model. Usual standard errors and tests are available, the graph shows pointwise 95% intervals, along with the p-value for a test of non-linearity. The bottom row compares the predicted 4 year death rate from the reference model to the predicted response from the logistic. In this case we have quantiles of the error, absolute deviation from the $y = x$ line, summary statistics suggested in Austin [**?**].

Yet to do: Find a compact way to summarize the nonlinear over the censoring distributions and the PH vs AFT models. A quick look shows that the censoring matters here too.

The third set of lines in Figures 1–4 show results for this counting process approach, and we see that it matches the expectations: the results are not affected by censoring patterns 1–3, nor by the underlying model type, as long as the model is correct.

Figures 3 and 4 highlight two aspects of the fit. First, the Cox model fit to the validation data set is not, in this case, a correct model for the data, and the result is no longer immune to changes in censoring pattern between random, a, and b. Second, in the no censoring case we see that the regression slope is seriously underestimated; this is a systematic bias due to fitting a proportional hazards model to data that is not PH. This latter effect is ameliorated by censoring: the non-proportionality effect of the log-logistic is diminished by a reduced time range.

If the validation data set is sufficiently large, an alternative is to fit a more flexible survival model to the validation data, and thereby avoid the more severe effects shown in this last case. As always, we will face the trade-offs between flexibility, efficiency and potential bias.

Looking at the last 5 lines of Figures 1 and 2 we see that this method is also unbiased, for all but the last censoring pattern; no method is immune to that bias. We also notice that the spread of values is tighter than others; this method appears to be more efficient than the counting process approach. The reason for this last, however, is that the prior three approaches use only the event information up to time τ, 4 years in this example, whereas the modeling approach commonly is coded to use all the events, both those before and after τ.

5.2 Risk score for amyloidosis

The amyloidosis example is based on a validation study that used follow-up from 1005 amyloidosis subjects from 12/2003 through 8/2015, from a single institution, to evaluate 4 different prediction models from the literature: published in 2004, 2012, 2013, and 2015 [**?**]. Each of these four models is based on a simple counting score. The 2004 model for instance uses 3 groups based on serum cardiac troponin $< .035$ μg/L (cTnT) and N-terminal pro-brain natriuretic peptide < 332 ng/L (NT-proBNP); subjects have 0, 1 or 2 markers above threshold. The 2012 and 2015 models define 4 risk groups, the 2013 model 5 groups. The latter two specifically build on the 2004

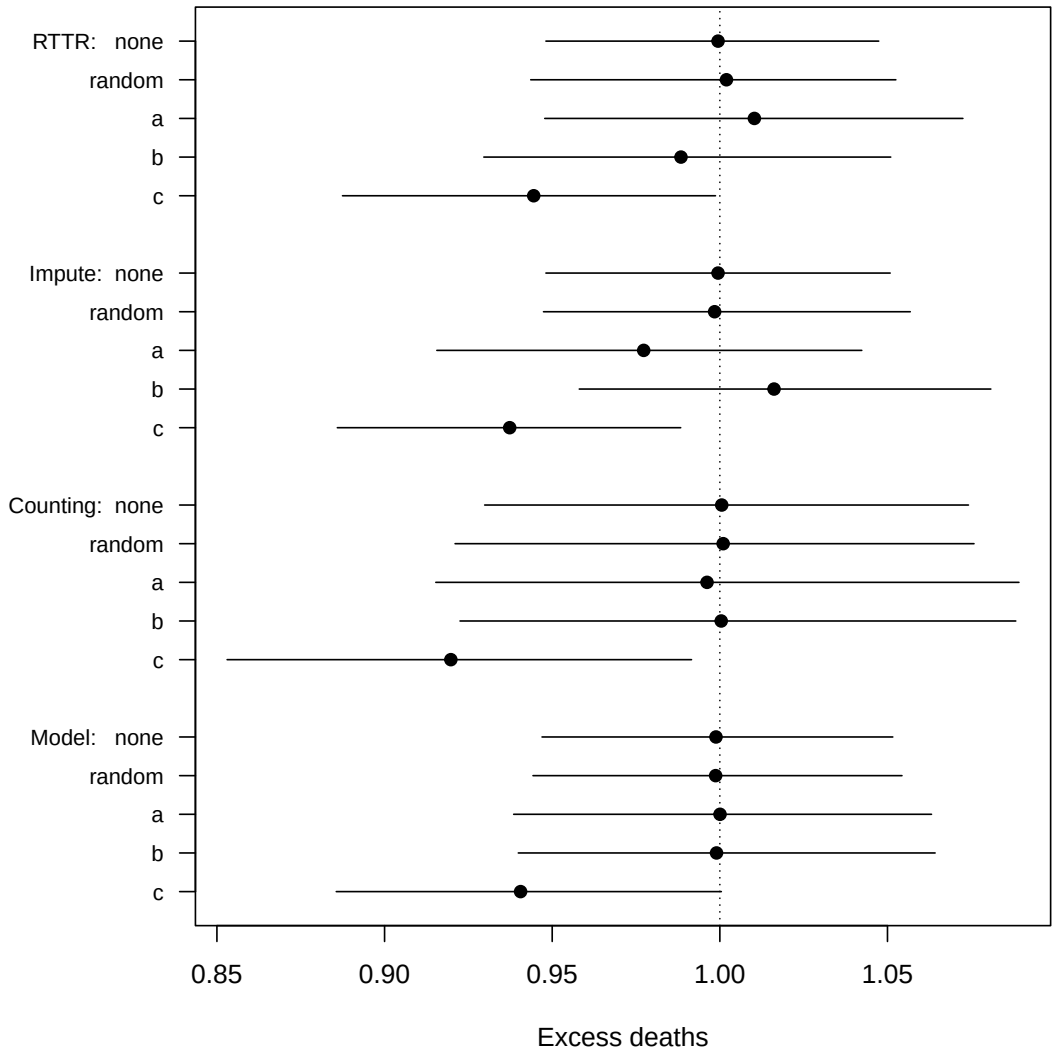

Figure 1: Results for four different methods when the true data comes from a proportional hazards model, with various censoring patterns. Each line spans the 10th and 90th percentiles of simulation results, the median is marked as a point.

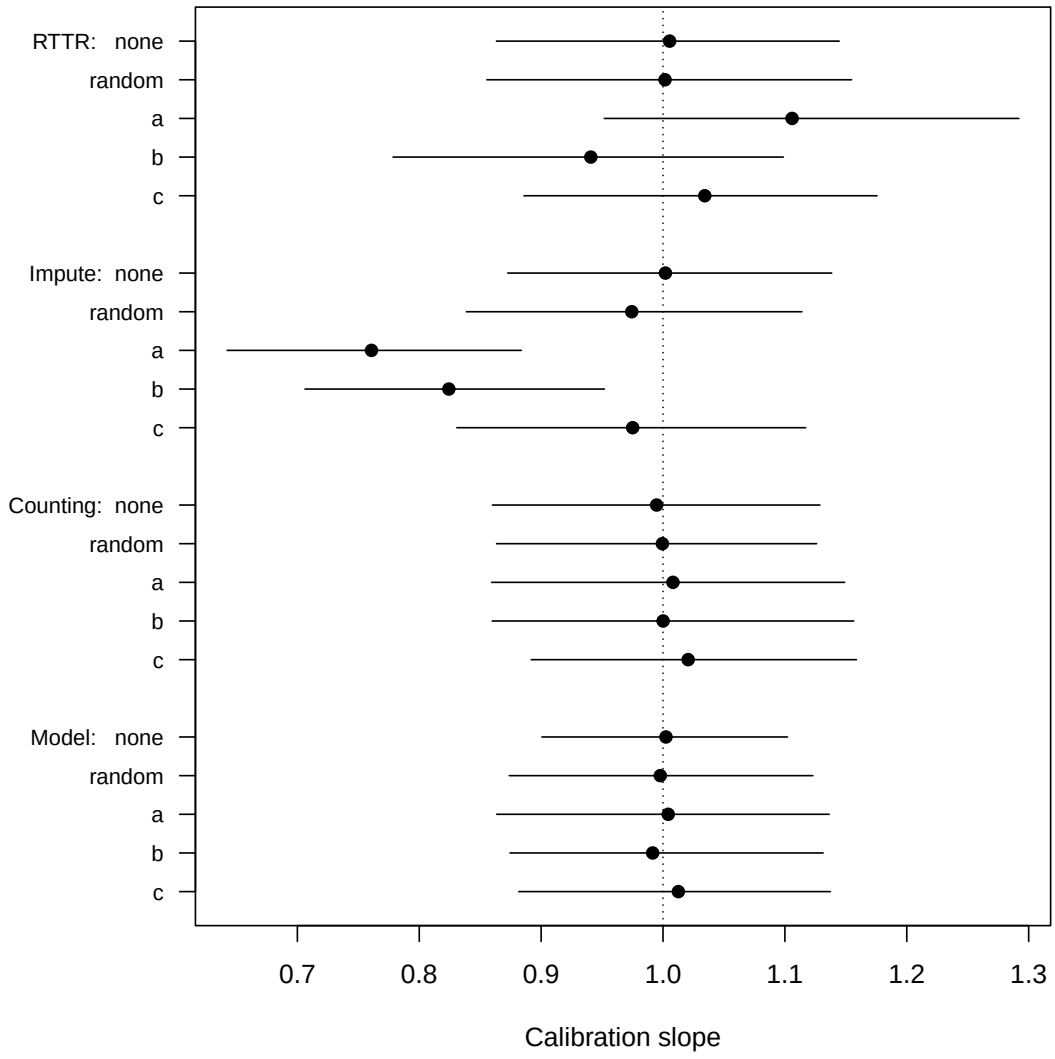

Figure 2: Results for four different methods when the true data comes from a proportional hazards model, with various censoring patterns. Each line spans the 10th and 90th percentiles of simulation results, the median is marked as a point.

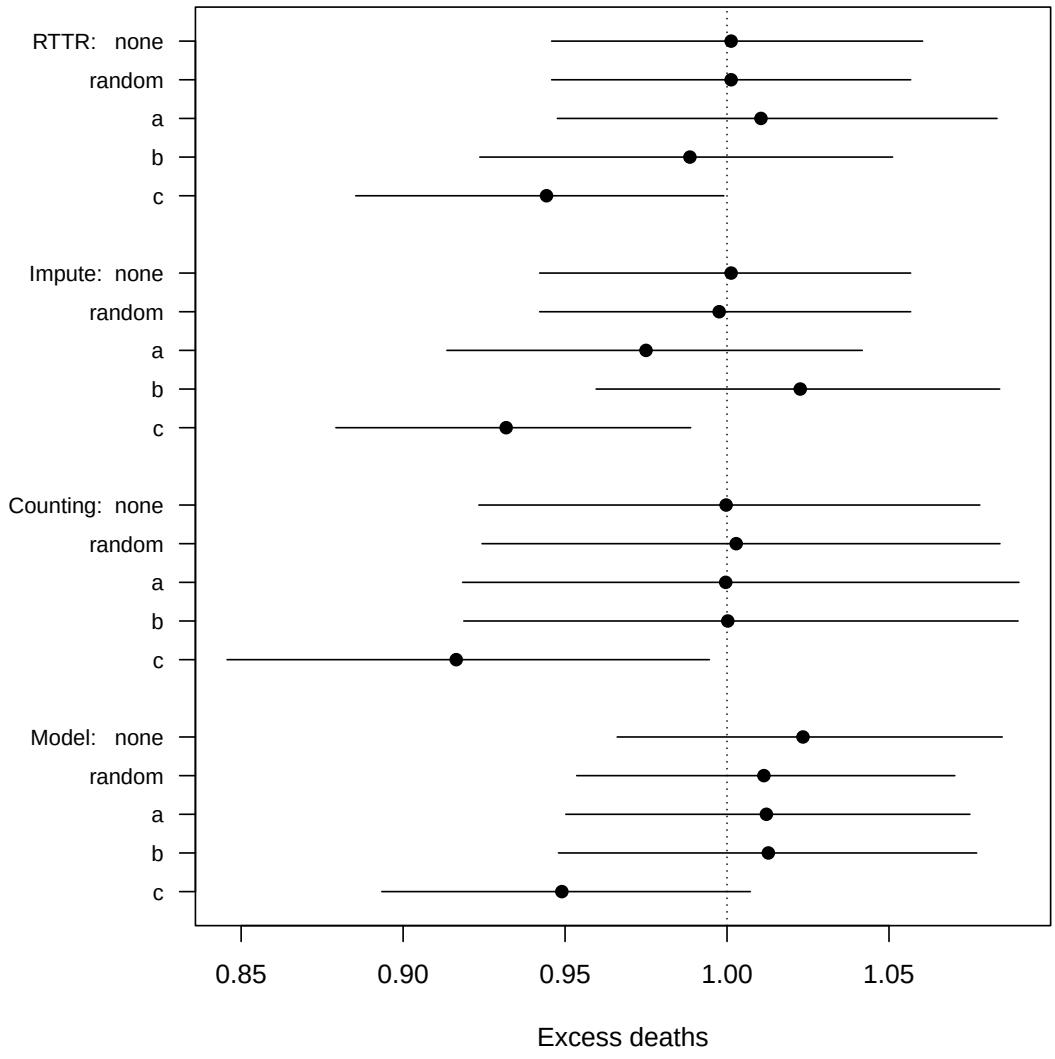

Figure 3: Results for four different methods when the true data comes from a log-logistic AFT model, with various censoring patterns. Each line spans the 10th and 90th percentiles of simulation results, the median is marked as a point.

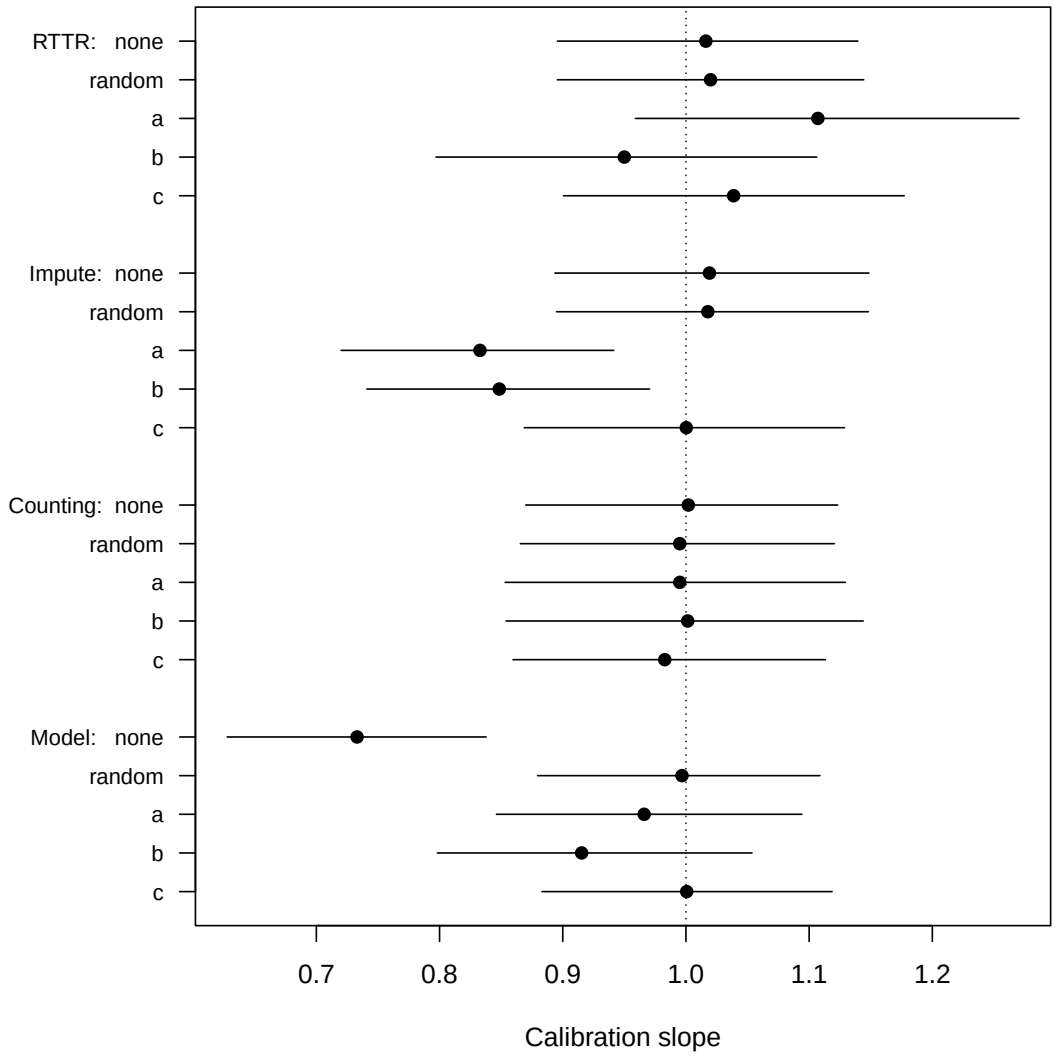

Figure 4: Results for four different methods when the true data comes from a log-logistic AFT model, with various censoring patterns. Each line spans the 10th and 90th percentiles of simulation results, the median is marked as a point.

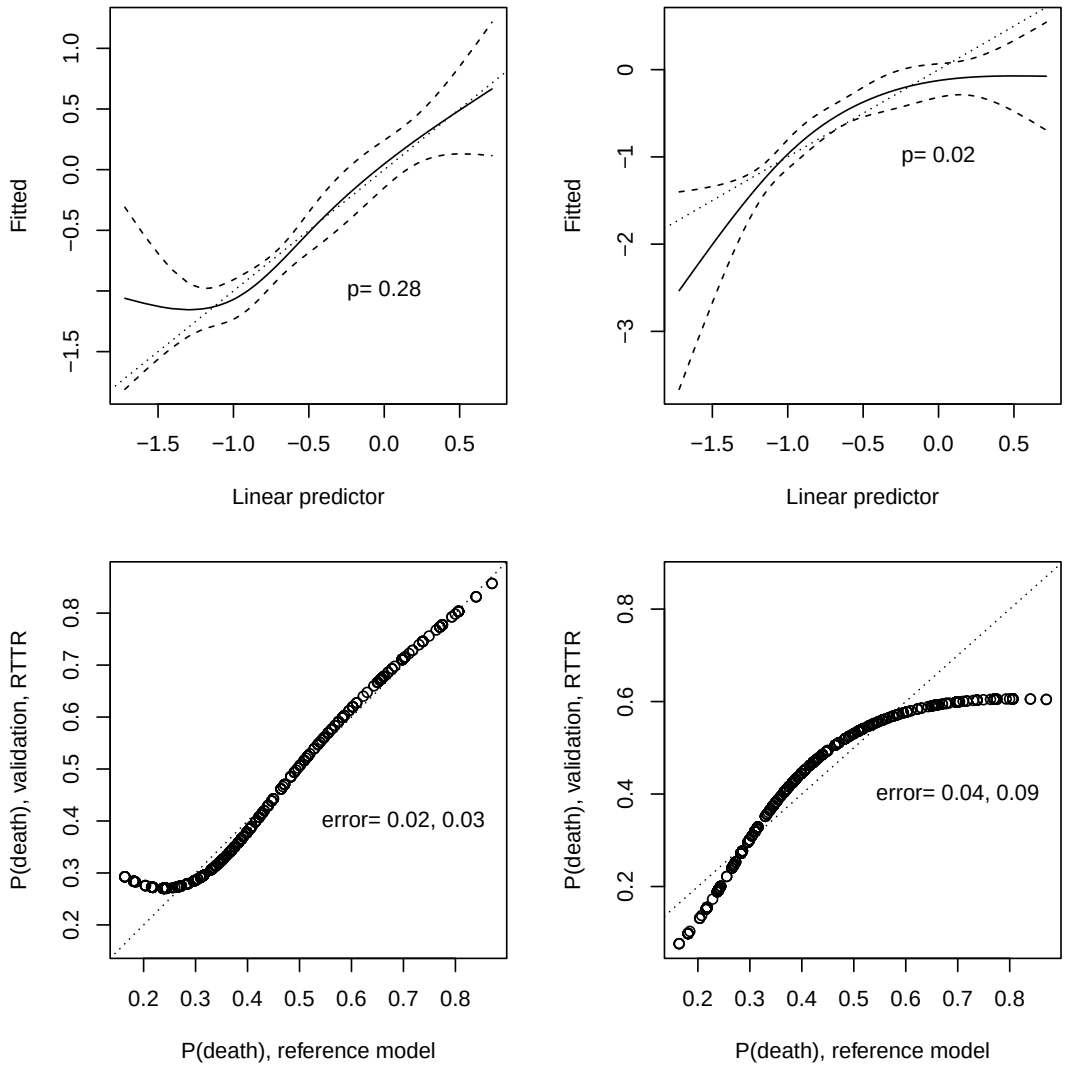

Figure 5: Spline fits for a case where the reference model is correct (left) and when there is a non-linear relationship between the predicted response of the reference model and the true risk (right). The top panels show standard logistic regression plots for the fitted spline, along with a p-value for a test for non-linearity. The bottom panels compare the predicted probability of death in 4 years to the predictions from the model. They include the 50th and 90th percentiles of the difference between the curve and the $y = x$ line. All panels include the $y = x$ reference as a dotted line.

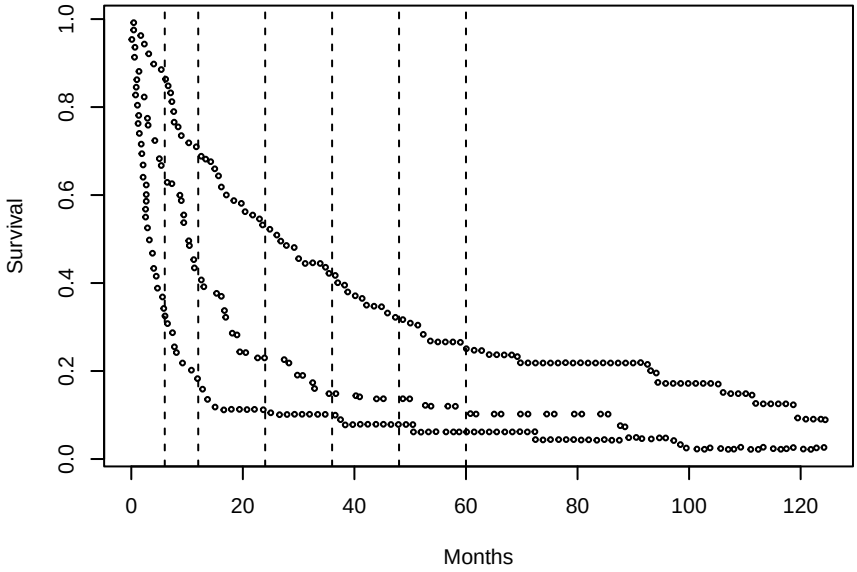

Figure 6: Digitized figure from the 2004 study.

model by using the same criteria for risk groups 0 and 1, then more finely dividing the highest risk subset. Predictions for the 1005 validation subjects were obtained by first digitizing the published survival curves from the four studies, then reading off values at 6, 12, 24, 36, 48 and 60 months. Where curves overlap or nearly overlap some judgement is required. Figure 6 shows the digitized data for the 2004 study.

We will declare 5 year survival as our clinical target, given the severity of the disease. Of the 1005 validation subjects 516 (51%) die before 5 years, 183 (18%) are censored before 5 years and 306 have 5+ years of follow-up. Table 1 shows the concordance values for the four prediction models, along with a simplistic 'baseline' model that uses only the number of involved organs (1–4) for each subject. We see that in terms of discrimination, there is little to choose between the four models, all of which are better than a simplistic prediction that uses only the number of involved organs. Different risk set weights have the most impact when there are few subjects remaining at risk, setting $\tau = 5$ makes the choice of weight largely irrelevant. The AUCROC is based on a discretized outcome. It addresses a slightly different question and is systematically somewhat larger, but also with a larger variance.

Formal comparisons between the values on any given line of the table can be computed, by making use of an infinitesimal jackknife variance. Using tau= 5 years, the concordance for organ count is significantly smaller than all others ($Z > 7.5$) and the Mayo 2004 concordance is smaller than the two Euro 2013 and 2015 models ($Z > 6.6$). [Compute and show this in the companion document? I'm not convinced that p-values are relevant here.]

Although the models are very close to equal in discrimination, there are large differences in predicted absolute risk. The underlying data is shown in table 2, which shows the number of subjects in each subgroup, for each of the risk models, along with predicted death rates, and observed death rates at 5 years for each subset in the validation data. The 2004 model was

	Organs	2004	2012	2013	2015
All time, Harrell	0.59	0.69	0.71	0.72	0.72
All time, Uno	0.56	0.67	0.68	0.69	0.69
0-5 years, Harrell	0.59	0.70	0.71	0.72	0.72
0-5 years, Uno	0.59	0.69	0.70	0.72	0.71
5 year AUROC	0.71	0.79	0.76	0.76	0.81
Cor with 2015	0.18	0.94	0.72	0.97	1.00

Table 1: Concordance between 5 different predictive models and the validation outcome, using either the Harrell or Uno weighting of risk sets, or an AUROC at 5 years. The Harrell and Uno C values have a standard error of 0.01, the AUROC values an se of .19–.24. The last line shows correlation between other C values and the 2015 model, for line 1.

	2004			2012				2015			
	0	1	2	0	1	2	3	I	II	IIIa	IIIb
n (validation)	199	380	426	227	217	246	315	199	380	235	191
6 year death, predicted	0.74	0.90	0.94	0.46	0.66	0.89	0.98	0.00	0.44	0.89	0.74
6 year death, KM	0.03	0.11	0.41	0.03	0.06	0.22	0.46	0.03	0.11	0.30	0.55

Table 2: Counts in the validation data set, predicted death rate at 6 years from each model, and Kaplan-Meier estimates in the validation data. (The 2013 paper's curve stop short of 5 years.)

widely adopted for risk stratification; the 2013 and 2015 publications suggest improvements that further subdivide only the highest stage; i.e., stage I of the 2015 paper is identical stage 0 of the 2004 model, yet the predicted 5 year death rate drops from 74% to 0. The large difference reflects the fact that the time period from 2004 to 2015 saw major gains in the treatment of amyloidosis and subsequent survival. The relative utility of the 2004 categorization remains, but absolute predictions are not useful.

For a simple categorical prediction such as this, the problem of censoring in the validation data set has a very simple solution, i.e., a per-subgroup Kaplan-Meier. The more sophisticated models can still be used, examples are displayed in the companion document.

5.3 Rotterdam and GBSG studies

A validation example is set forth in Royston and Alman [**?**] that uses 2982 primary breast cancers patients whose records were included in the Rotterdam tumor bank data to build an initial model, and 626 subjects in a 1984-1989 trial from the German Breast Study group as the validation data set. Importantly, the data was made available for others and has been used in many subsequent papers. Figure 7 shows the overall survival curves for both cohorts.

Initial exploration showed that menopausal status, tumor size, tumor grade, and the number of positive lymph nodes are important variables for a multivariate survival model in the Rotterdam data. The number of lymph nodes is a continuous variable with a wide range of 0–34, a spline fit to the data shows a fairly linear increase in risk up to 10 nodes, and none thereafter. Table xxx shows the fitted Cox model.

As will always be the case, the data sets do not exactly match. The most obvious difference is shorter follow-up for the GBSG study. Another is that the Rotterdam data contains follow-up

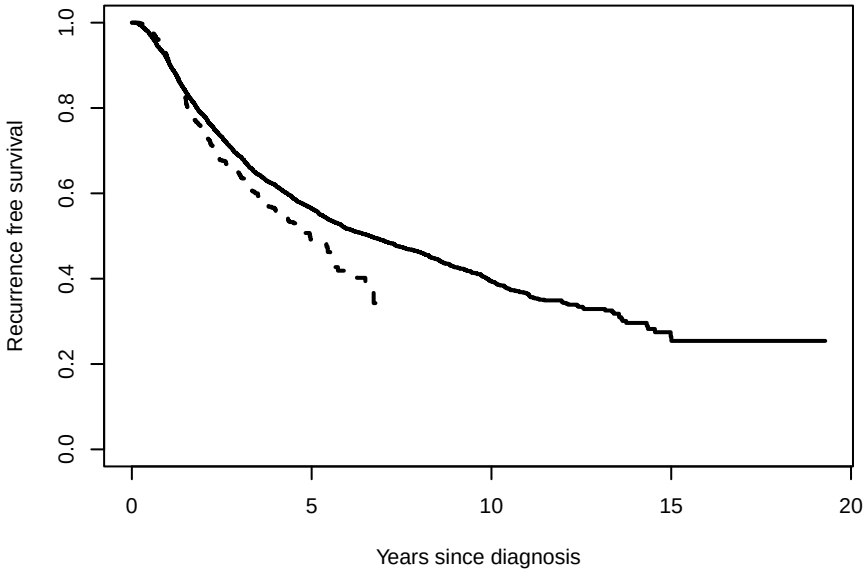

Figure 7: Overall survival for the Rotterdam (solid) and GBSG (dashed) participants.

to death and to recurrence as separate variables, while the GBSG data set has a single endpoint of recurrence free survival (the earlier of recurrence or death). The GBSG data has continuous tumor size while the Rotterdam tumor data is categorical: 0-20, 20-50, and 50+ cm; so we create a categorical version in the GBSG data. With respect to nodes, 48% of the Rotterdam patients are node-negative (0 affected lymph nodes), while none of the GBSG subjects are. Rotterdam subjects have grade 2 or 3, GBSG grades 1-3. None of these are fatal to the comparison, but should always be borne in mind. For instance, if the step from grade 1 to grade 2 is much larger, biologically, then the step from 2 to 3, then extrapolation of Rotterdam results to grade 1 GBSG subjects might be questionable.

```
> # replace this with a table
> print(rfit, digits=2)
Call:
coxph(formula = Surv(ryear, rfs) ~ meno + size + grade + pmin(nodes,
    10), data = rott2)

                  coef exp(coef) se(coef)    z      p
meno            0.1096    1.1158   0.0497  2.2   0.03
size20-50       0.3044    1.3558   0.0547  5.6  3e-08
size>50         0.5246    1.6897   0.0823  6.4  2e-10
grade           0.3153    1.3707   0.0597  5.3  1e-07
pmin(nodes, 10) 0.1300    1.1389   0.0072 18.2 <2e-16

Likelihood ratio test=568  on 5 df, p=<2e-16
```

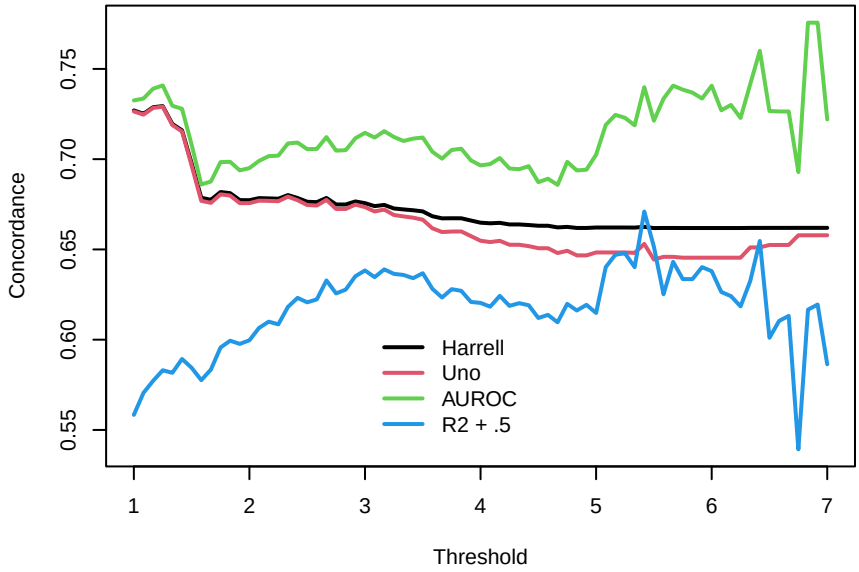

Figure 8: Concordance of the Rotterdam prediction for the GBSG data set, as a function of the threshold τ that is chosen, along with the estimated AUROC and R^2 for that time point based on IPC weights.

```
n= 2982, number of events= 1713
```

The baseline or reference fit, using the Rotterdam data, will use all the subject's data. If there is evidence for non-proportional hazards (and there is), then one can reasonably argue that if validation will be focused at τ years, then reference fit restricted to the first τ years after enrollment would be more likely to replicate. Our rationale for using all the data is that first, in the common case where the reference fit was done first, with later external validation based in the published report of that fit, the original authors would not know that a "4 year" model will later be desired, and secondly that many published fits ignore issues with PH. [This latter comment may be too pessimistic.]

Figure 8 shows multiple indices of discrimination as a function of the chosen threshold τ. The C statistics have a median standard error of .017 and the AUROC a median of .024, i.e., the visible differences between these three are not particularly noteworthy, though the patterns are interesting. The AUROC does become unstable at the far right, but this is not surprising. At year 7 there are only 3 subjects remaining in the risk set, each with an IPC weight of 78.4; 'validating' seven year survival when so few have sufficient follow-up is unrealistic. R^2 for the categorized 0/1 response, which is a scaled Brier score, follows a similar pattern as the AUROC.

Figure 9 gives an assessment of potential non-linear effects, often referred to as a 'moderate' assessments of calibration; the models use the linear predictor from the reference model as the single predictor in a spline term. The effect is so close to linear that a separate test for an overall slope of 1 ('weak' calibration) is not really necessary in this case. We can also plot curves on probability scale as shown in Figure 10, for the modeling method, as suggested in Austin et al

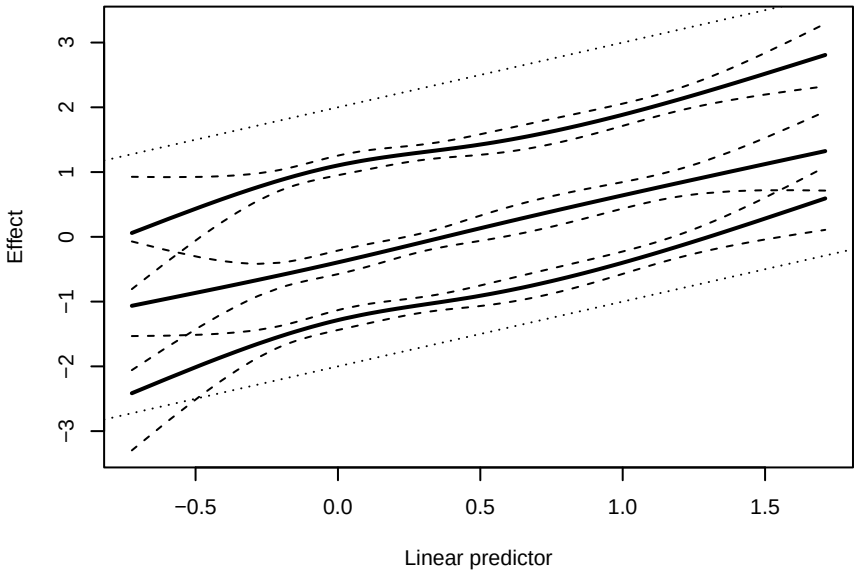

Figure 9: Estimated relationship between the linear predictor, based on the reference model, and the outcomes in the validation data set. The dotted lines have slope=1 and are for visual reference, dashed lines are pointwise 95% confidence intervals, solid are based on a natural spline with 3 degrees of freedom. Top line: counting process approach, offset by +1. Middle line: RTTR approach. Bottom line: modeling approach, offset by -1.

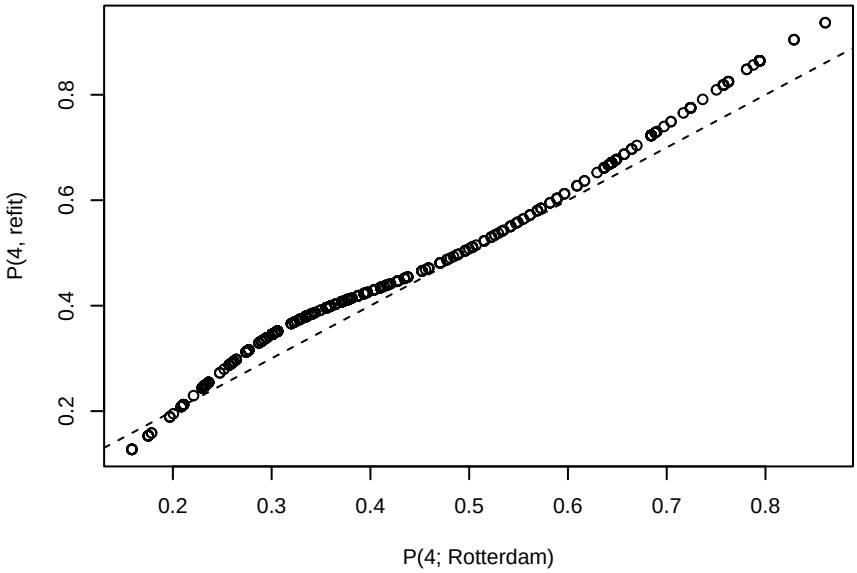

Figure 10: Predicted values of the probability of recurrence or death within four years under the reference model, fit to the Rotterdam data, versus the predicted value from a Cox model fit directly to the validation data.

[**?**]. They label x an y axes in this figure as "predicted" and "observed", but we would argue that the latter label is dishonest. It is actually a plot of the prediction under the reference model (x) versus prediction under a model fit directly to the validation data. (A good idea but a bad label).

As an example, GBSG validation of the Rotterdam model is perhaps too good, and thus does not teach us very much. Since the GBSG data is from a clinical trial the censoring is almost entirely administrative and thus independent of survival, and calibration is near perfect. In such a scenario all the methods work identically.

[Note: I don't understand why the counting process shows about a death rate of .98 of expected, and the modeling approach 1.07 of expected.]

5.4 PBC

5.5 PBC

The `pbc` data set in R contains information for 312 subjects with primary biliary cirhossis (PBC) who were enrolled in a clinical trial of D-penicilline versus placebo, along with another 106 who declined to be randomized but agreed to baseline laboratory tests and yearly follow-up. As a simple test, fit a Cox model to the randomized subjects, and validate it on the passive subjects. It is well known that the set of subjects who agree to enrollment in a clinical trial is often a select subset.

Figure 11 shows the distribution of risk scores for the two cohorts, along with overall survival

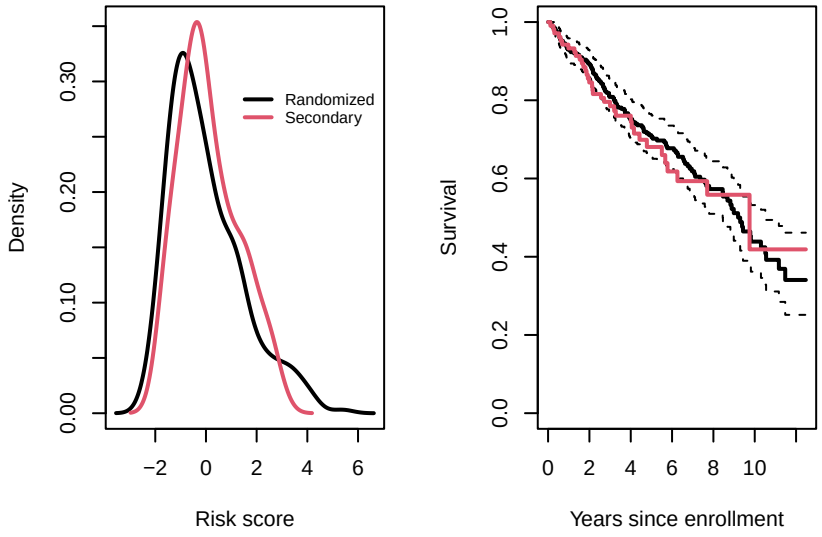

Figure 11: Fitted risk scores $X\beta$ for the randomized subjects and predicted risk scores for the others (left panel), along with KM curves for the two subsets.

curves. The mean risk score for the reference data has been set to 0 by default, predicted risk score for the validation data has a mean of 0.12.

This is as dull as the GBSG data set I just complained about.

More interseting, if I can dig up the old data, might be the placebo and arms of the UDCA study in PBC. It is an interesting story: UDCA works, and the changes in liver lab tests that it induced were such that the the PBC risk score, applied using updated data, is surprisingly accurate. That is, the effect of UDCA was well captured by the liver function tests used in the PBC risk score. The risk score is a surrogate endpoint.

6 Monoclonal Gammopathy

There are 2 data sets in the package that contain follow-up on patients with MGUS, and earlier time frame with 241, and a later one with 1384. There are a few patients in common (30) which I will ignore for the moment.

```
> fit2 <-  coxph(Surv(futime, death) ~ age + sex + pmin(hgb, 14), mgus)
> fit2
Call:
coxph(formula = Surv(futime, death) ~ age + sex + pmin(hgb, 14),
    data = mgus)

                    coef exp(coef)  se(coef)      z        p
```

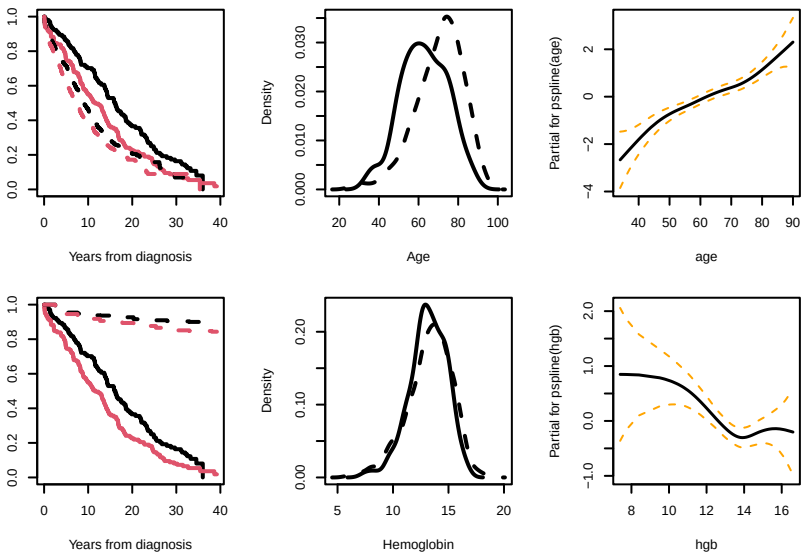

Figure 12: Survival and covariates for the MGUS data.

```
age            0.066271  1.068517  0.007047  9.404  < 2e-16
sexmale        0.337844  1.401922  0.140780  2.400   0.0164
pmin(hgb, 14) -0.223754  0.799512  0.051679 -4.330 1.49e-05

Likelihood ratio test=120.2  on 3 df, p=< 2.2e-16
n= 240, number of events= 224
   (1 observation deleted due to missingness)
```