The Stata Journal (2014)
14, Number 4, pp. 738–755

Tools for checking calibration of a Cox model in external validation: Approach based on individual event probabilities

Patrick Royston
MRC Clinical Trials Unit at University College London
London, UK
j.royston@ucl.ac.uk

Abstract. The Cox proportional hazards model has been used extensively in medicine over the last 40 years. A popular application is to develop a multivariable prediction model, often a prognostic model to predict the clinical outcome of patients with a particular disorder from "baseline" factors measured at some initial time point. For such a model to be useful in practice, it must be "validated"; that is, it must perform satisfactorily in an external sample of patients independent of the sample on which the model was originally developed. One key aspect of performance is calibration, which is the accuracy of prediction, particularly of survival (or equivalently, failure or event) probabilities at any time after the time origin. We believe systematic evaluation of the calibration of a Cox model has been largely ignored in the literature. In this article, we suggest an approach to assessing calibration using individual event probabilities estimated at different time points. We exemplify the method by detailed analysis of two datasets in the disease primary biliary cirrhosis; the datasets comprise a derivation and a validation dataset. We describe a new command, `stcoxcal`, that performs the necessary calculations. Results for `stcoxcal` can be displayed graphically, which makes it easier for users to picture calibration (or lack thereof) according to follow-up time.

Keywords: st0357, stcoxcal, Cox proportional hazards model, multivariable model, prognostic factors, external validation, calibration, survival probabilities

1 Introduction

Multivariable survival models are used in medicine, particularly as the basis of prognostic models in clinical and research practice and as risk models for population screening. Given certain "baseline" factors measured at some appropriate initial time point (denoted by $t = 0$), the models are used to predict the future clinical outcome of individuals with, for example, a particular condition such as cancer or heart disease or those at risk of such. Risk models have several applications, including selection of persons at high risk of needing preventive therapy, stratification of risk in clinical trials and audit studies, and personalized prediction of disease outcome.

Models developed for patients in a given sample may predict well within that dataset but fail to "generalize" (predict) well on samples from other patient populations. Assessing generalizability is the key component of external model validation. In statistical terms, external validation involves checking that outcome predictions from a model developed on a "derivation" sample are sufficiently accurate in an independent "validation" sample. See Altman and Royston (2000) and Altman et al. (2009) for a general background on model validation.

It is common to distinguish two aspects when "validating" a model: discrimination and calibration. Discrimination means the ability of a model to distinguish between outcomes of patients with different risks. Calibration means the accuracy of prediction, particularly of survival (or equivalently, failure or event) probabilities at any time after $t = 0$. The literature has paid much attention to measures of discrimination; popular examples include the work of Harrell et al. (1982) on the c index of concordance and that of Royston and Sauerbrei (2004) on the D statistic. (See Choodari-Oskooei, Royston, and Parmar [2012] for a detailed comparison between several approaches.) In contrast, little has been published on assessing calibration of models for time-to-event data; Harrell (2001) describes techniques based on the bootstrap for internal validation. Here we focus on tools for assessing calibration in external data.

We suggest an approach to assessing calibration of a Cox proportional hazards model using individual event probabilities at different time points. This may be seen as an extension of recent work (Royston and Altman 2013; Royston Forthcoming) addressing calibration in aggregates of patients at similar risk of an event (that is, in risk groups). The tools may also be used to check the calibration of a model on the same data it was developed on, that is, on the derivation dataset. See Royston and Altman (2013) for further considerations in validating a published Cox model.

The structure of this article is as follows. We first outline the framework for our survival modeling. We describe two datasets from the disease primary biliary cirrhosis (PBC) as well as the Cox model we are using as a running example. We discuss our motivation and approach to calibration in a more familiar logistic regression context, and we extend it to the Cox proportional hazards model framework. We present illustrative analyses using the PBC data. We then explain `stcoxcal`, a new tool that implements the analyses and graphs. We finish with some closing comments.

2 Proportional hazards models

Suppose we have a vector of explanatory variables $\mathbf{x} = (x_1, \ldots, x_k)$. A Cox proportional hazards model with the parameter vector $\boldsymbol{\beta}$ incorporates multiplicative effects of $\mathbf{x}$ on the baseline hazard function and is usually written as

$$h(t; \mathbf{x}) = h_0(t) \exp(\mathbf{x}\boldsymbol{\beta}) \tag{1}$$

where $h(t;\mathbf{x})$ is the hazard function, $h_0(t) = h(t;\mathbf{0})$ is the baseline hazard function, and t is the follow-up time. If we integrate (1), we obtain the cumulative hazard function, $H(t;\mathbf{x}) = \int_0^t h(u;\mathbf{x})du$. Taking logarithms, we get

$$\ln H(t;\mathbf{x}) = \ln H_0(t) + \mathbf{x}\boldsymbol{\beta} \tag{2}$$

Now let's write $\ln H(t;\mathbf{x}) = g\{S(t;\mathbf{x})\}$, where $S(t;\mathbf{x})$ is the survival function and $g(u) = \ln(-\ln u)$ is the "link function". Then we can write (2) as

$$g\{S(t;\mathbf{x})\} = g\{S_0(t)\} + \mathbf{x}\boldsymbol{\beta} \tag{3}$$

In the Cox model (1), the baseline hazard function and, hence, the baseline survival function in (3), $S_0(t) = S(t;\mathbf{0})$, are unspecified and are not estimated as part of the model.

3 Example datasets

3.1 Description and proportional hazards model

To illustrate, we assemble prognostic variables in common across two datasets relating to the disease PBC (usually known as cirrhosis of the liver). The first dataset on which we derived a model was used by Fleming and Harrington (1991) to exemplify certain aspects of survival analysis. The data comprise survival or censoring times of $n = 418$ patients (161 deaths) with PBC, 312 of whom entered a randomized controlled trial and the remaining 126 participated in a cohort study. Six prognostic factors with complete data were available for analysis.

The second dataset, which we used as a validation sample, comes from a randomized controlled trial of 248 patients with PBC (Christensen et al. 1985). After removing 41 cases (17%) with missing values or no patient follow-up, we had data on 207 patients (105 deaths) for analysis.

Three covariates were recorded in both datasets: age, bilirubin, and albumin. We applied `mfp` with Cox regression (the `stcox` command) to build the following proportional hazards model in the derivation dataset:

$$h(t;\mathbf{x}) = h_0(t)\exp\{0.04085 \times \text{age} + 0.9405 \times \ln(\text{bilirubin}) - 0.09852 \times \text{albumin}\}$$

The predictive ability is high for a survival model: Harrell's $c = 0.824$ and Royston and Sauerbrei's $D = 2.27$, for which $R^2_D = 55\%$. Corresponding values in the validation dataset when applying the prognostic index (PI) $\mathbf{x}\widehat{\boldsymbol{\beta}}$ from the derivation dataset are Harrell's $c = 0.785$ and Royston and Sauerbrei's $D = 1.89$, for which $R^2_D = 46\%$. There appears to be some reduction in the discrimination of the model in the validation dataset.

The two datasets were combined for analysis. A binary variable `val` was created, taking the value zero in the derivation dataset and one in the validation dataset.

3.2 Preliminary analysis

Figure 1a shows Kaplan–Meier survival curves in the derivation and validation datasets. Survival (unadjusted for covariates) is clearly worse in the validation dataset. Figure 1b shows estimates of the baseline survival curve in each dataset, which are computed by centering the PI on its mean in the derivation dataset and offsetting the PI in Cox models separately in each dataset. Although adjusting the PI clearly brought the curves closer together (and, incidentally, altered their shape), the lower survival in the validation dataset persists. Figure 1b suggests that the model is imperfectly calibrated in the validation dataset, with a tendency to underpredict event probabilities. By fitting the following Cox model to the combined dataset,

```
. stcox val xb
```

we find that the adjusted hazard ratio for `val` is 1.38 [95% confidence interval (CI) 1.08, 1.77], which shows an increased adjusted hazard for `val` = 1, as expected. The unadjusted hazard ratio for `val` is 1.70. Next we investigate the calibration in more detail.

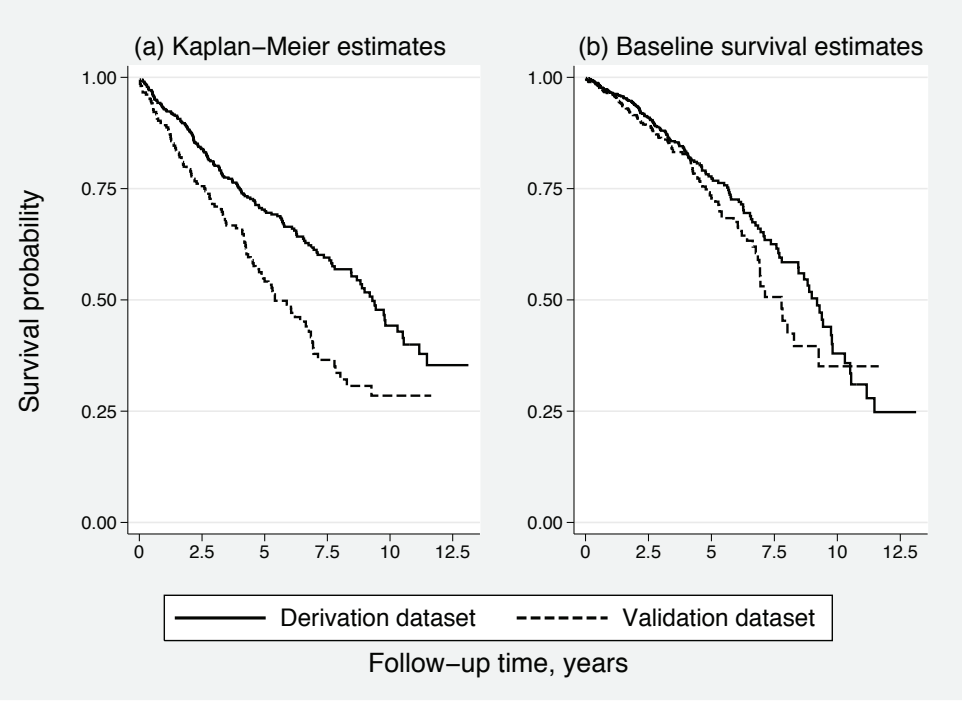

Figure 1. PBC datasets. a) Kaplan–Meier survival curves; b) baseline survival curves according to the PI centered on 0 in the derivation dataset.

We clarify the Stata-related details for obtaining the curves in figure 1b as follows. The three covariates are `x1` (age), `x2` (log bilirubin), and `x3` (albumin).

```
stcox x1 x2 x3 if val==0
predict xb, xb // note: predicts for all observations, including val=1
summarize xb if val==0
replace xb = xb - r(mean)
stcox if val==0, offset(xb)
predict s0, basesurv
stcox if val==1, offset(xb)
predict s1, basesurv
line s0 s1 _t, sort
```

We centered `xb` to ensure that the baseline distribution function is meaningful. Thus `xb` $= 0$ represents a patient in the derivation dataset at "average risk" of dying.

4 Assessing calibration of logistic regression models

Let $F(t;\mathbf{x}) = 1-S(t;\mathbf{x})$ be the failure (event) probability, that is, the chance of an event occurring in the interval $(0,t)$ for an individual with covariate vector $\mathbf{x}$. To motivate what follows, we first consider a logistic regression model. Now t plays no role, so the event probability, $F(\mathbf{x})$, is a function of only the PI, $\boldsymbol{x\beta}$, and the baseline log odds of an event, $\beta_0 = \text{logit}\{F(\mathbf{0})\}$. Assessing model calibration means comparing the observed event probabilities with those predicted by the model. The observed event probability for an individual is taken as 1 if the individual experiences an event (outcome $Y = 1$) and 0 otherwise (outcome $Y = 0$). We write the PI as PI$= \widehat{\beta}_0 + \mathbf{x}\widehat{\boldsymbol{\beta}}$. The predicted event probability is $\widehat{F}(\mathbf{x}) = \text{logit}^{-1}(\text{PI}) = \{1 + \exp(-\text{PI})\}^{-1}$. An auxiliary logistic regression model, which is linear in the PI (Miller, Hui, and Tierney 1991), may be used to check agreement between observed and predicted probabilities.

$$\text{logit}\{\Pr(Y = 1)\} = \gamma_0 + \gamma_1 \text{PI} \tag{4}$$

If model (4) is fit to the same dataset as that used to estimate $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$, the estimates of γ_0 and γ_1 are identically 0 and 1, respectively, which is of no help. However, (4) may be used to investigate external validation when $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ are estimates from a published report or other suitable source.

Consider the simplified auxiliary model

$$\text{logit}\{\Pr(Y = 1)\} = \gamma_0 + \text{PI} \tag{5}$$

that is, with γ_1 constrained to 1, (4) with the PI offset from the linear predictor. The intercept γ_0 in (5) assesses calibration "in the large" (Harrell 2001) because it shifts the entire distribution function $F(Y)$ by γ_0 on the logit scale.

We can quantify miscalibration easily by applying three hypothesis tests based on (4) and (5). For calibration in the large, we fit (5), estimate γ_0, and test $\gamma_0 = 0$. To check the regression on the PI (essentially, discrimination), we fit (4) and test $\gamma_1 = 1$. To perform an overall test of calibration, we use a joint test of $(\gamma_0, \gamma_1) = (0, 1)$ with 2

degrees of freedom. If we are concerned about type 1 error, a conservative approach is to perform the joint test first, then proceed to the separate tests of γ_0 and γ_1 only if the result of the joint test is significant. This is a closed-test procedure that maintains the familywise error rate.

Before we return to survival models, we note that calibration error in the validation dataset may be a more complex function of the PI than a straight line on the logistic scale. Thus (4) and (5) may be misspecified. A recommended graphical adjunct is to plot a scatterplot smooth of Y or of residuals $Y - \widehat{F}(\mathbf{x})$ on $\widehat{F}(\mathbf{x})$, together with pointwise CIs. This can reveal subtle miscalibration. It can also be used as a graphical check of calibration on the derivation dataset.

5 Assessing calibration of Cox regression models

5.1 The baseline distribution function

The principle of checking calibration by comparing observed and predicted event probabilities can also be used with Cox models. Calibration in the survival context is intrinsically time dependent. It may be assessed overall and at several suitable values of t up to the maximum event time.

This raises an important issue for the Cox model. In external validation, we wish to evaluate predicted event probabilities, $\widehat{F}(t;\mathbf{x})$ for some t, in an independent dataset. To do this, we need to "export" the baseline distribution function, $F_0(t) = 1 - S(t;\mathbf{0})$, for relevant values of t from the derivation data to the validation data. However, the Cox model does not provide an estimate of the baseline distribution function. A simple solution to this is to model the baseline distribution function in the derivation dataset using a suitable class of approximating models. Often an adequate solution is to model the log baseline cumulative-hazard function, $\ln H_0(t) = \ln\{-\ln S_0(t)\}$, as a second-degree fractional polynomial (FP2) in t (Royston and Altman 2013; Royston Forthcoming). Despite the assumptions of linear regression analysis not being met, the function to be approximated is very smooth, and it is satisfactory to estimate the parameters of the FP2 model $E\{\ln H_0(t)\} = \delta_0 + \delta_1 t^{p_1} + \delta_2 t^{p_2}$ by ordinary least squares. p_1 and p_2 may be estimated using the `fracpoly` or (in Stata 13.0 and above) the `fp` command. See Royston and Altman (1997) for further examples of the usefulness of FP functions for approximation of smooth functions.

The ordinary least-squares regression comprises two stages. First, after fitting the Cox model to the derivation dataset using `stcox`, we use the command `predict` *varname*, `basechazard` to estimate the baseline log cumulative-hazard function in the derivation dataset. We then regress *varname* on t as previously described. The resulting FP2 function can be used to predict $\ln H_0(t)$ out of sample in the validation dataset. The out-of-sample prediction step cannot readily be done without the intermediate regression analysis.

5.2 Overall calibration

In principle, we can investigate the calibration for external validation by using Cox regression on the PI in the validation sample. As with logistic regression, we are interested in the parameters γ_0 and γ_1 in this global setting. However, the Cox model has no intercept, so we cannot estimate γ_0. Furthermore, by regressing on the PI in the validation dataset, we are reestimating the baseline distribution function. For a strict assessment of calibration, we wish to avoid such reestimation. We want to know whether the entire model (baseline included) fit on the derivation dataset still predicts accurately in the validation dataset.

We take a different approach to obtain a calibration model with a linear predictor of the form $\gamma_0 + \gamma_1$PI. Instead of the PI, the covariate in the model for the ith patient at time t is the estimated log cumulative-hazard function, $\ln \widehat{H}(t; \mathbf{x}_i) = \ln[-\ln\{1 - \widehat{F}(t; \mathbf{x}_i)\}]$, which is the complementary log-log transformation of the predicted event probability, $\widehat{F}(t; \mathbf{x}_i)$. We obtain the predicted event probability by "importing" the baseline distribution function with an FP2 function estimated in the derivation data and applied out of sample to the validation data, as previously described. We have

$$\widehat{F}(t; \mathbf{x}_i) = 1 - \widehat{S}_0(t)^{\exp(\mathrm{PI}_i)}$$

where

$$\ln\left\{-\ln \widehat{S}_0(t)\right\} = \hat{\delta}_0 + \hat{\delta}_1 t^{p_1} + \hat{\delta}_2 t^{p_2}$$

We thus have "expected" event probabilities at time t at the individual level. How do we derive corresponding "observed" event probabilities? Pohar Perme and Andersen (2008) and Andersen and Pohar Perme (2010) help with this. Given a sample of n individuals and a time point t within the observed follow-up times, the method of pseudo-observations (which we call "pseudovalues") provides values $\widetilde{F}_1(t), \ldots, \widetilde{F}_n(t)$, which are unbiased estimates of $F_1(t), \ldots, F_n(t)$. Right-censoring is taken into account. Parner and Andersen (2010) elegantly implemented the method in Stata as the `stpsurv` command. Note that the values $\widetilde{F}_1(t), \ldots, \widetilde{F}_n(t)$ are jackknife quantities and individually do not resemble recognizable event probabilities. For example, they are not necessarily confined to $(0, 1)$ and may even be negative or exceed 1. Their key property is their unbiasedness in expectation.

For any reasonable value of t, the Cox model is perfectly calibrated on the validation dataset if the following property holds:

$$E\left\{\widetilde{F}_i(t)\right\} = F(t; \mathbf{x}_i)$$

Under these conditions, a generalized linear model (GLM) with responses $\widetilde{F}_1(t), \ldots, \widetilde{F}_n(t)$, linear predictor $\gamma_0 + \gamma_1 \ln \widehat{H}(t; \mathbf{x}_i)$, and complementary log-log link function $\ln\{-\ln(1 - x)\}$ should fit the validation data well.

5.3 Testing regimen: Single time point

The GLM with responses $\widetilde{F}_i(t)$ and linear predictor $\gamma_0 + \gamma_1 \ln \widehat{H}(t; \mathbf{x}_i)$ supports three tests at time t:

1. Intercept test. Constraining $\gamma_1 = 1$, we wish to know whether $\widehat{\gamma}_0$ is consistent with 0. If it is not consistent with 0, we have a calibration error sometimes known as "miscalibration in the large" (Harrell 2001). The (adjusted) event rate in the validation data differs from that in the derivation data.

2. Slope test. Test of $\gamma_1 = 1$ with γ_0 estimated. For a correctly calibrated model, the estimate $\widehat{\gamma}_1$ should be consistent with 1. If the test p-value is significant, the calibration failed. Because $\gamma_1 = 0$ implies a complete lack of model discrimination, the case $\widehat{\gamma}_1 < 1$ may also indicate reduced discrimination in the validation dataset.

3. Joint test. The GLM can also furnish a joint test of $(\gamma_0, \gamma_1) = (0, 1)$ with 2 degrees of freedom. This examines the overall evidence for (linear) miscalibration.

One approach yielding a closed-test procedure is to perform the joint test (test 3) first and, if it is significant, to proceed to the intercept and slope tests. The last two tests should provide more information on the nature of the miscalibration.

5.4 Example

We illustrate the three tests for investigating calibration in the PBC validation dataset. We chose the time point $t = 7$ years. Table 1 shows Wald tests from the GLM reported by the `stcoxcal` command.

Table 1. Tests of calibration in the PBC dataset at $t = 7$ years

Test	Wald χ^2	Degrees of freedom	p-value
1. Intercept	6.42	1	0.011
2. Slope	0.03	1	0.85
3. Joint	6.81	2	0.033

The estimate of the intercept is $\widehat{\gamma}_0 = 0.44$ (standard error 0.17), suggesting that the mortality rate (relative hazard) at 7 years is a factor of about $\exp(0.44) = 1.55$ higher in the validation dataset. There is no evidence that $\gamma_1 \neq 1$ ($p = 0.85$). The joint test with 2 degrees of freedom provides some evidence ($p = 0.033$) that $(\gamma_0, \gamma_1) \neq (0, 1)$. Because $\widehat{\gamma}_1$ is consistent with 1, the miscalibration seems to be entirely in the large. Notice that the joint test is less significant and has lower power in this situation. There is almost no contribution from $\gamma_1 \neq 1$ toward the 2 degrees of freedom χ^2 statistic.

The `stcoxcal` command and its output with the results given in table 1 and with the various parameter estimates from the GLMs are shown below.

```
. use pbc, clear
(PBC data, 3 sources)
. stcox x1 x2 x3 if !val
  (output omitted)
. predict xb, xb
. stcoxcal xb, val(val) times(7) test
                                  (Std. Err. adjusted for 207 clusters in _id)
------------------------------------------------------------------------------
          _f |      Coef.   Semirobust Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1._times |   .4416587    .174341     2.53   0.011     .0999568    .7833607
      _clogF |          1  (offset)
------------------------------------------------------------------------------

[Test 1: intercepts (gamma0) = 0 with slope (gamma1) constrained to 1]
 ( 1)  1._times = 0

           chi2(  1) =    6.42
         Prob > chi2 =    0.0113
                                  (Std. Err. adjusted for 207 clusters in _id)
------------------------------------------------------------------------------
          _f |      Coef.   Semirobust Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _clogF |   .9585466   .2217061     4.32   0.000     .5240107    1.393083
    1._times |   .4192495   .2173986     1.93   0.054    -.0068439    .8453429
------------------------------------------------------------------------------

[Test 2: slope (gamma1) = 1 with constants (gamma0) estimated]
 ( 1)  _clogF = 1

           chi2(  1) =    0.03
         Prob > chi2 =    0.8517
[Test 3: joint test of slope (gamma1) = 1 and all constants (gamma0) = 0]
 ( 1)  1._times = 0
 ( 2)  _clogF = 1

           chi2(  2) =    6.81
         Prob > chi2 =    0.0333
```

The graph provided by `stcoxcal` is shown in figure 2. The dashed line shows the (rather jagged) running-line smooth of the pseudovalues at $t = 7$ years, estimated by `running`. The smoothed event probabilities are higher than predicted at all values of the predicted probability, demonstrating miscalibration in the large.

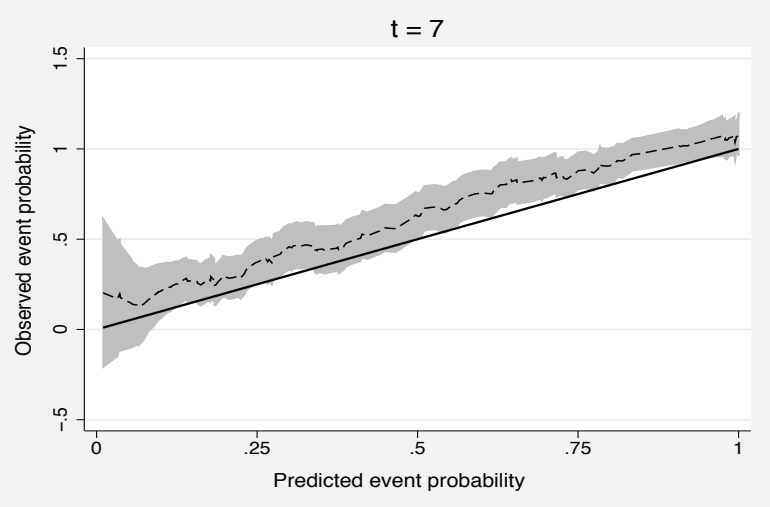

Figure 2. Smoothed pseudovalues (dashed lines) with pointwise 95% CI plotted against predicted event probabilities for the PBC data at $t = 7$ years. The solid line is the line of identity, denoting perfect calibration. Some miscalibration in the large is evident, with underprediction of event probabilities in the validation dataset.

5.5 Testing regimen: Multiple time points

In practice, we do not wish to limit calibration assessment to one time point. Rather, we want to assess it at several time points $t_1, \ldots, t_m$ spanning the follow-up period. We use `stpsurv` to estimate the m pseudovalues for every patient. The calibration analysis can now provide m models $\gamma_{0j} + \gamma_{1j} \ln \widehat{H}(t_j; \mathbf{x})$ $(j = 1, \ldots, m)$, with 1 linear predictor for each time point.

Such models may be fit for all m time points simultaneously after reshaping the data to "long" format structured according to the m replicates. The `glm` command with `link(cloglog)` is used to fit models to the mn pseudovalues $\widetilde{F}_i(t_j)$ on $\ln \widehat{H}(t_j; \mathbf{x}_i)$ $(i = 1, \ldots, n;\ j = 1, \ldots, m)$. The sandwich estimator is used for variance estimation of regression coefficients (Andersen and Pohar Perme 2010).

We do not want to fit such a complex model with as many as m linear predictors, so we must simplify it. We start by investigating miscalibration in the large. We assume distinct intercept parameters $\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0m}$ and a single slope γ_1, and we constrain $\gamma_1 = 1$. We implement this by using `glm` with t as a factor variable and by applying the `offset()` option. Test 1, the test for zero intercepts, tests $\gamma_{01} = \gamma_{02} = \cdots = \gamma_{0m} = 0$ using a Wald test with m degrees of freedom. Test 2, the slope test, is similar to the $m = 1$ case: we assume $\gamma_{11} = \gamma_{12} = \cdots = \gamma_{1m}$ and again include t as a factor variable. We test whether the average value of γ_1 equals 1. Test 3, the joint test, is again performed with $\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0m}$ and γ_1 fitted, and it has $m + 1$ degrees of freedom.

We can expand the model by testing for an interaction between $\ln \widehat{H}(t_j; \mathbf{x})$ and the m time points. This allows us to investigate whether $\gamma_{11} = \gamma_{12} = \cdots = \gamma_{1m}$, that is, whether γ_1 changes over time. A statistically significant test of the interaction suggests that miscalibration varies. For example, a Cox model in data with long-term follow-up could predict accurately in early follow-up but fail later by losing discrimination (reduction in γ_1) or by changing miscalibration in the large (change in γ_0).

The test of interaction has $m - 1$ degrees of freedom, with the null hypothesis being that $\gamma_{11} = \gamma_{12} = \cdots = \gamma_{1m}$. This is calculated using the user-written command `stcoxcal` with the `test` option described in section 6.

5.6 Example

We extend the example with $t_1 = 7$ years by considering yearly intervals up to 9 years; that is, $t_j = j$ $(j = 1, \ldots, 9)$. The results of the four calibration tests are shown in table 2.

Table 2. Tests of calibration in the PBC dataset over nine years of follow-up

Test	Wald χ^2	Degrees of freedom	p-value
1. Intercept	16.49	9	0.057
2. Slope	0.42	1	0.52
3. Joint	18.00	10	0.055
4. Interaction	8.13	8	0.42

Test 1, the test of intercepts, shows some evidence against all intercepts being 0, but it is borderline ($p = 0.057$). Test 2, the slope test of $\gamma_1 = 1$, is nonsignificant ($p = 0.52$). Considered over all 9 time points, there is no evidence that $\gamma_1 \neq 1$. A borderline result is also obtained for the joint test. The interaction test does not suggest that γ_1 varies over time.

5.7 Plotting smoothed pseudovalues across time

Under perfect calibration, the vector $\widehat{F}(t; \mathbf{x})$ should accurately describe the expected pseudovalues across patients at any appropriate time t. Because miscalibration may be time dependent, plotting a combined graph with the chosen values of t may help to clarify the pattern of miscalibration over time.

The results for the PBC data are shown in figure 3.

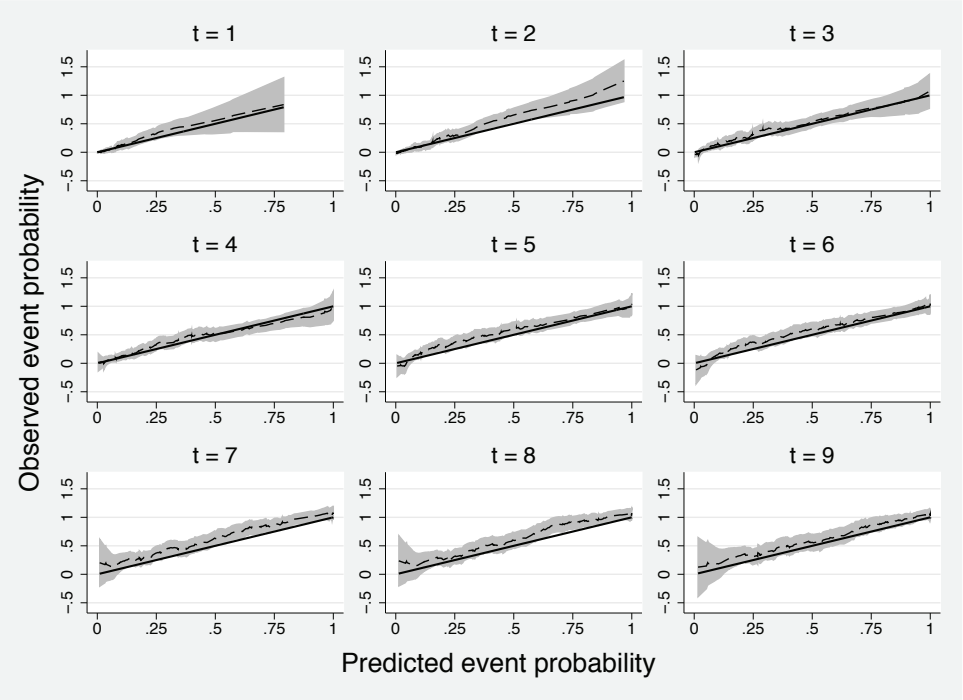

Figure 3. Smoothed pseudovalues (dashed lines) with pointwise 95% CI plotted against predicted event probabilities for the PBC data. Times represent each year over nine years of follow-up. The solid line is the line of identity, denoting perfect calibration.

The underestimation of event probabilities in the validation dataset is visible at most of the nine time points.

For a more detailed view, it is sometimes helpful to smooth and plot the residuals, that is, the pseudovalues minus the predicted event probabilities. Under perfect calibration, there should be no important biases, trends, or patterns among the smoothed residuals. The corresponding plot requires the `residuals` option of `stcoxcal`. Figure 4 shows this type of plot for the PBC data. Essentially, the same message emerges as from figure 3.

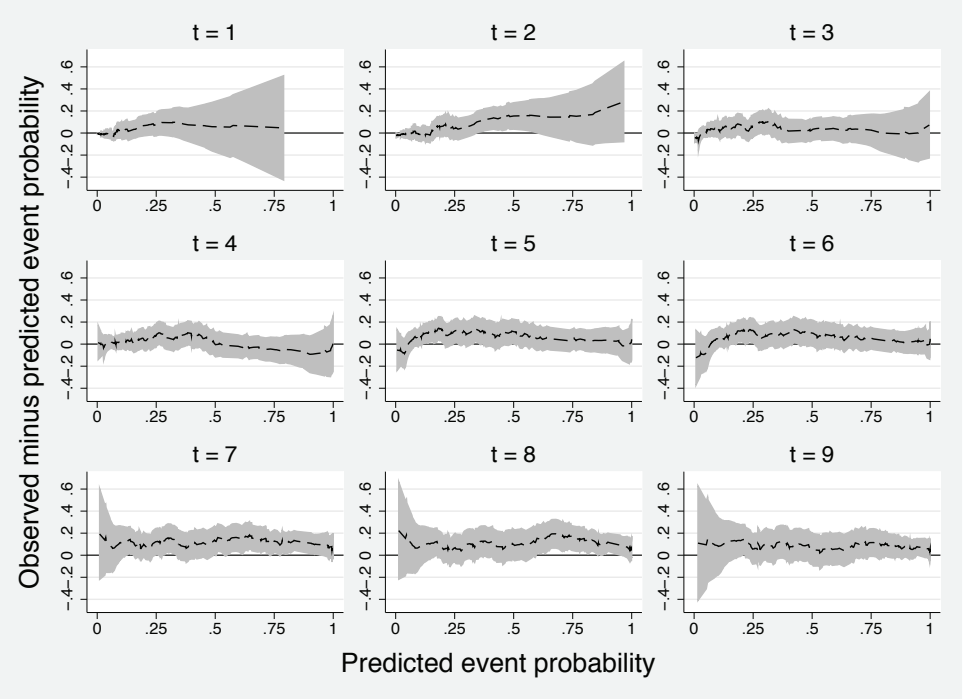

Figure 4. Smoothed pseudovalue residuals (dashed lines) with pointwise 95% CI plotted against predicted event probabilities for the PBC data. The horizontal line at $y = 0$ denotes perfect calibration.The plot uses the `residuals` option of `stcoxcal`.

6 Implementation

The methods, calculations, and graphs described above are implemented with the command `stcoxcal`.

6.1 Syntax

`stcoxcal` *xbetavar* [*if*] [*in*], `times(`*numlist*`)` [`nograph` `residuals` `saving(`*filename*`)` `test` `trend` `val(`*varname*`)` *graph_twoway_options* *running_options*]

You must have `running` and `stpsurv` installed before using `stcoxcal`. `running` is available from the Statistical Software Components archive (see [R] **ssc**). `stpsurv` can be installed from the *Stata Journal* archive with the command `net install st0202, from(http://www.stata-journal.com/software/sj10-3)`.

6.2 Description

`stcoxcal` is a tool for examining the (possibly time-dependent) calibration of a Cox model whose linear predictor ("prognostic index") is supplied in *xbetavar*.

A "well-calibrated" model is one that accurately predicts survival or event probabilities at all relevant follow-up times. A model that includes covariates whose effects change (for example, dwindle) over time is unlikely to be well calibrated. Such a model will give a more or less biased prediction of survival probabilities. `stcoxcal` is designed to detect and display the lack of calibration graphically. It also includes tests of good calibration and of time-dependent trends of miscalibration.

By default, `stcoxcal` examines calibration of a model on its "own" dataset. With the `val()` option, `stcoxcal` can be used to examine model calibration in an independent dataset (that is, for external validation).

6.3 Options

`times(`*numlist*`)` lists times at which model calibration is to be assessed. `times()` is required.

`nograph` suppresses the production of calibration plots.

`residuals` plots the smoothed residuals (difference between observed and predicted event probabilities) against the predicted event probabilities. The default is to plot smoothed observed against predicted event probabilities.

`saving(`*filename*`)` saves five variables in the validation dataset to file *filename*:

`_id`	observation number in the original data
`_times`	integer scores (levels) 1, 2, ... of times specified in `times()`
`_f`	pseudovalues for event probabilities
`_F`	event probabilities predicted from a Cox model
`_clogF`	complementary log-log transformation of `_F`

These variables can be used by an expert to create plots and to further analyze model calibration. The data are held in long format, with a complete set of values for each level of `_times`.

`test` tests whether the slope (on the log cumulative-hazard scale) of the regression of pseudovalues for event probabilities on predicted event probabilities over all time points in `times()` equals one. A nonsignificant p-value suggests good overall calibration, sometimes called "calibration in the large". `test` also tests the interaction between the slopes and the times specified in `times()`. A significant p-value suggests that calibration changes over time. Typically, calibration declines as follow-up time increases.

`trend` tests whether the slope (on the log cumulative-hazard scale) of the regression of pseudovalues for event probabilities on predicted event probabilities over all time

points in `times()` equals one (same as for `test`). `trend` also tests the linear interaction between the slopes and the integer scores for the times specified in `times()`. This may be more powerful than the interaction test provided by `test`.

`val(`*varname*`)` is for use in external validation. *varname* is a binary variable coded zero to define the "model derivation" dataset and any other nonmissing value to define the "model validation" dataset. Predictions of event probabilities at different times from the derivation dataset are made in the validation dataset via the linear predictor and a smoothed version of the baseline cumulative-hazard function in the derivation dataset. Royston and Altman (2013) call this "strict" calibration.

graph_twoway_options are options of `graph twoway`. These may be used to customize the appearance of the calibration plots.

running_options are options of `running`. These may be used to customize the smoothing of pseudovalues. The most relevant option is likely to be `span(#)`. See `help running` for further information.

6.4 Remarks

Note that `stcoxcal` computes the baseline survival and cumulative hazard functions internally. As a preliminary, `stcoxcal` centers the prognostic index supplied in *xbetavar* on zero. If `val(`*varname*`)` is provided, the mean of *xbetavar* in the subset defined by *varname* = 0 is subtracted from all values of *xbetavar*. Otherwise, centering takes place over the estimation sample. Next, a Cox model is fit with no covariates and with *xbetavar* offset from the linear predictor. Again, this is done either in the *varname* = 0 subset or in the estimation sample. Finally, the baseline cumulative hazard function is predicted and smoothed for use with the calibration method described in section 5.

Because *xbetavar* or indeed the original covariates are not refitted to the validation data, `stcoxcal` can be used in "partial validation" mode. The prognostic index is created from a derivation model fit elsewhere and imported for application in the available validation dataset. Validation is partial because the baseline cumulative hazard and survival functions are estimated by `stcoxcal` on the validation data, whereas *xbetavar* is calculated by the user on the validation data from regression coefficients estimated externally. Although imperfect, partial validation nevertheless allows a useful evaluation of the predictive accuracy of a predefined model when the baseline distribution function is (perforce) tailored to the validation data.

6.5 Examples

```
webuse brcancer, clear
stset rectime, failure(censrec) scale(365.24)
fp generate x1^(-2 -0.5)
fp generate x6^(0.5), scale
stcox x1_1 x1_2 x4a x4b x5e x6_1 hormon
predict xb, xb
stcoxcal xb, times(1(1)6) test
stcoxcal xb, times(1(1)6) trend

set seed 3143
generate byte random_half = (runiform() < 0.5)
stcox x1_1 x1_2 x4a x4b x5e x6_1 hormon if random_half==0
predict xb2, xb
stcoxcal xb2, val(random_half) times(1(1)6) test

stcox x1 x4a x4b x5 x6 hormon
predict xb3, xb
stcoxcal xb3, times(1(1)6) test
```

6.6 Stored results

`stcoxcal` stores the following in `r()`:

Scalars

`r(gamma1)`	estimate of γ_1 with γ_0 estimate
`r(gamma1_se)`	Std. Err. of `gamma1`
`r(P0)`	p-value for test 1, of $\gamma_0 = 0$ given $\gamma_1 = 1$
`r(P1)`	p-value for test 2, of $\gamma_1 = 1$ with γ_0 estimated
`r(P01)`	p-value for test 3, joint test of $(\gamma_0, \gamma_1) = (0, 1)$
`r(Pint)`	p-value for test 4, of interaction of γ_1 with time

Macros

`r(fp_pwrs)`	powers of `_t` in FP2 model for $\ln H_0(t)$

7 Comments

Currently, `stcoxcal` works only with "plain" Cox models, that is, Cox models without stratification factors (the `strata()` option in `stcox`) and without time-dependent regression coefficients (`tvc()` not allowed) or time-varying covariates.

The user can choose the time points at which to assess calibration. Results will somewhat vary with different choices, so it may be advisable to perform a sensitivity analysis. For example, when one uses the $m = 5$ time points of 2, 4, 6, 8, and 10 years with the PBC data, test 1 is significant at $p = 0.040$ and test 3 at $p = 0.045$. This result should change the interpretation only slightly. The estimates of γ_{0j} $(j = 1, \ldots, 5)$ with $\gamma_1 = 1$ are 0.32, 0.06, 0.28, 0.40, and 0.24, again suggesting some miscalibration in the large.

If a formula for the baseline survival function calculated on the derivation data were provided, `stcoxcal` could be slightly extended to handle validation of a published model even without the raw derivation data. This would require investigators proposing a Cox

model to publish an FP-based expression for the baseline log cumulative-hazard function. Unfortunately, in practice, the baseline survival function or a suitable transformation of it is never reported, so implementing the extension is not currently worthwhile. Nevertheless, as described in section 6.4, partial validation of an externally derived prognostic index is an option in such cases.

Calibration of Cox proportional hazards models has been substantially neglected in the literature. We hope that our suggestions as well as the new command `stcoxcal` for assessing calibration of these models will help analysts with external validation of time-to-event models.

8 References

Altman, D. G., and P. Royston. 2000. What do we mean by validating a prognostic model? *Statistics in Medicine* 19: 453–473.

Altman, D. G., Y. Vergouwe, P. Royston, and K. G. M. Moons. 2009. Prognosis and prognostic research: Validating a prognostic model. *British Medical Journal* 338: b605.

Andersen, P. K., and M. Pohar Perme. 2010. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 19: 71–99.

Choodari-Oskooei, B., P. Royston, and M. K. B. Parmar. 2012. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine* 31: 2627–2643.

Christensen, E., J. Neuberger, J. Crowe, D. G. Altman, H. Popper, B. Portmann, D. Doniach, L. Ranek, N. Tygstrup, and R. Williams. 1985. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: Final results of an international trial. *Gastroenterology* 89: 1084–1091.

Fleming, T. R., and D. P. Harrington. 1991. *Counting Processes and Survival Analysis*. New York: Wiley.

Harrell, F. E., Jr. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Harrell, F. E., Jr., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247: 2543–2546.

Miller, M. E., S. L. Hui, and W. M. Tierney. 1991. Validation techniques for logistic regression models. *Statistics in Medicine* 10: 1213–1226.

Parner, E. T., and P. K. Andersen. 2010. Regression analysis of censored data using pseudo-observations. *Stata Journal* 10: 408–422.

Pohar Perme, M., and P. K. Andersen. 2008. Checking hazard regression models using pseudo-observations. *Statistics in Medicine* 27: 5309–5328.

Royston, P. Forthcoming. Tools for checking calibration of a Cox model in external validation: Graphical approach based on risk groups. *Stata Journal*.

Royston, P., and D. G. Altman. 1997. Approximating statistical functions by using fractional polynomial regression. *Journal of the Royal Statistical Society, Series D* 46: 411–422.

———. 2013. External validation of a Cox prognostic model: Principles and methods. *BMC Medical Research Methodology* 13: 33.

Royston, P., and W. Sauerbrei. 2004. A new measure of prognostic separation in survival data. *Statistics in Medicine* 23: 723–748.

About the author

Patrick Royston is a medical statistician with more than 30 years of experience, with a strong interest in biostatistical methods and in statistical computing and algorithms. He works largely in methodological issues in the design and analysis of clinical trials and observational studies. He is currently focusing on alternative outcome measures in trials with a time-to-event outcome; on problems of model building and validation with survival data, including prognostic factor studies and treatment-covariate interactions; on parametric modeling of survival data; and on novel clinical trial designs.