Biometrics 68, 380–387
June 2012
DOI: 10.1111/j.1541-0420.2012.01753.x

Choice of Prognostic Estimators in Joint Models by Estimating Differences of Expected Conditional Kullback–Leibler Risks

Daniel Commenges,* Benoit Liquet, and Cécile Proust-Lima*****

University of Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique,
F-33000 Bordeaux, France

**email:* daniel.commenges@isped.u-bordeaux2.fr

***email:* benoit.liquet@isped.u-bordeaux2.fr

****email:* cecile.proust@isped.u-bordeaux2.fr

Summary. Prognostic estimators for a clinical event may use repeated measurements of markers in addition to fixed covariates. These measurements can be linked to the clinical event by joint models that involve latent features. When the objective is to choose between different prognosis estimators based on joint models, the conventional Akaike information criterion is not well adapted and decision should be based on predictive accuracy. We define an adapted risk function called expected prognostic cross-entropy. We define another risk function for the case of right-censored observations, the expected prognostic observed cross-entropy (EPOCE). These risks can be estimated by leave-one-out cross-validation, for which we give approximate formulas and asymptotic distributions. The approximated cross-validated estimator CVPOL_a of EPOCE is studied in simulation and applied to the comparison of several joint latent class models for prognosis of recurrence of prostate cancer using prostate-specific antigen measurements.

Key words: AIC; Cross-entropy; Estimator choice; Joint models; Kullback–Leibler risk; Likelihood cross-validation; Prognosis; Prostate cancer.

1. Introduction

In clinical practice, different therapeutical procedures may be proposed to subjects with different prognoses. Statistical models may help in this task. Most often the issue is to predict the occurrence of a clinical event using the information available at the time the prognosis is made. Conventional survival analysis tools should thus be useful for this aim. One difference, however, is that most often, in addition to fixed covariates, repeated measurements of markers are available, and these measurements carry valuable information for prognosis. One example is the prostate-specific antigen (PSA) which can be used for prognosis of clinical recurrence of prostate cancer. Following external beam radiation therapy (EBRT), PSA levels reach a nadir, and generally remain a certain time at low levels with possibly a very slow increase. A subsequent steeper rise of PSA is strongly associated with prostate cancer recurrence or death (Taylor, Yu, and Sandler, 2005; Proust-Lima et al., 2008). To more accurately guide clinical decision making, the complete posttreatment PSA trajectory after EBRT could be used for evaluating the risk of recurrence. Such a prognosis can be based on joint models that simultaneously describe the biomarker trajectory and the associated risk of event.

Joint models can be based on latent processes (Henderson, Diggle, and Dobson, 2000). In the case where the latent process is not degenerate, when for instance a Brownian motion is involved, the computation of the likelihood is challenging. Joint latent class models (JLCMs) have been proposed as an alternative requiring more manageable computations (Lin et al., 2002; Proust-Lima and Taylor, 2009). The principle of these models is to assume that subjects belong to different latent classes that are characterized by different distributions of both markers and events. Whatever their specification, joint models help us to build estimators of the prognosis distribution at time t, that is the distribution of the event of interest conditional on information available at time t; such estimators will be called "prognosis estimators."

When several prognosis estimators are available the question arises to choose between them. The issue is not to choose "the good model" but to choose the prognosis estimator with the best predictive ability. Although conventional criteria such as Akaike information criterion (AIC) (Akaike, 1973) or the Bayesian information criterion (BIC) (Schwarz, 1978) may be useful in conventional cases, they are not adapted for choosing prognosis estimators based on joint models; the same is true for conventional likelihood cross-validation (LCV) criterion. More specific criteria of predictive accuracy have been proposed in the standard survival context, in particular, by Schemper and Henderson (2000) and Gerds and Schumacher (2006), both proposing generalizations of the mean absolute error or of the mean squared error (also known as Brier score), although Heagerty and Zheng (2005) proposed time-dependent receiver operating characteristic curves. Predictive accuracy measures were also extended to assess dynamic

prognostic tools derived from joint models (Henderson, Diggle, and Dobson, 2002; Proust-Lima and Taylor, 2009; Schoop, Graf, and Schumacher, 2008). However, in presence of censoring most of these approaches require strong assumptions. Moreover, these approaches do not correct for overoptimism coming from the fact that the estimator uses the information incorporated into the criterion. Gerds and Schumacher (2007) proposed a bootstrap approach for correcting for overoptimism but this is not computationally feasible for joint models.

The aim of this article is to find a criterion for choosing between different prognosis estimators, which can be easily extended to the censored case. In the first place, we will define the (theoretical) risk functions based on information theory, the expected cross-entropy in the uncensored and in the censored case; then we have to find estimators of these risks. In contrast with previous works in the predictive accuracy context, it allows to select the prognosis estimator based on the whole distribution of the time of the event rather than a binary variable coding whether the event occurred in a given time window. One advantage of the approach is that in case of censoring the theoretical criterion remains easy to estimate, even if the models are misspecified, and this does not require estimating the censoring distribution. We correct for the overoptimism and provide an estimate of the variability of the risk estimator.

Section 2 develops the theory: Section 2.1 presents the context of prognosis, where the relationship between the predictive variables and the time of the event to be predicted is mediated through a latent element; Section 2.2 defines the relevant risk, the expected prognostic cross-entropy (EPCE). Section 3 presents estimators of EPCE and difference of EPCE in the case of uncensored observations. In the case of censored observations, another theoretical criterion is proposed, the expected prognostic observed cross-entropy (EPOCE) which is presented in Section 4. These theoretical risks can be estimated via cross-validation and an approximate formula is given. The asymptotic distributions of these risk estimators are given in Section 5. Two extensions are discussed in Section 6: a penalized likelihood approach for relaxing parametric assumptions, and an integrated criterion which allows us to propose an estimator based on a unique model for all the prediction times. Section 7 presents an illustration with real data of prostate cancer. Section 8 concludes. Supplementary Material contains proofs and a simulation study which allows assessing the properties of the estimators of the theoretical risks.

2. The Expected Cross-Entropy for Prognosis

2.1 *The Context of Prognosis*

Given the information available at a certain time we are interested in predicting the time T of occurrence of an event. In the simple prognosis situation, characteristics X of a subject are available at the time-origin and we would like to know the (true) distribution of T given X, which can be specified by the conditional density $f^*_{T|X}$. Technically, if a sample $\bar{\mathcal{O}}_n = (T_i, X_i, i = 1, \ldots, n)$ is available, estimating $f^*_{T|X}$ is a standard regression problem. Because T is a time, it is typical that some observations are right censored, that is, we observe $(\tilde{T}_i, \delta_i)$, where $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = 1_{T_i \leq C_i}$, C_i being a censoring variable. Estimation with right-censored observations has been thoroughly studied in survival analysis. The choice of estimators based on information theory has been promoted by Akaike (1973).

Estimation and choice of estimators for dynamic prognosis raise different issues. Here, we wish to make predictions not only at time zero but potentially at any time t. The information available for a particular subject at time t includes the survival event $\{T > t\}$, fixed covariates X, and repeated marker values up to t, $\bar{Z}_t$. All the density functions that we will use will be potentially conditioned on X; for the sake of notational simplicity this conditioning will be omitted in the following. We would like to know the true prognostic density $f^*_{T|\bar{Z}_t, T>t}$.

A simple way to solve the problem of estimating this predictive density by standard techniques is to change the time-origin, putting it in t, which apparently reduces the problem to the simple prognosis one, as done within the landmark analysis (Van Houwelingen, 2007). For estimation, one then uses only the subjects such that $\tilde{T}_i > t$. This approach however has several drawbacks. First, for large t only little information is available for estimation. Second, different subjects may have different numbers of marker measurements so that standard regression techniques cannot use the whole available information. Third, for different values of t the models are different, and may be mutually inconsistent.

A more efficient approach is to use a joint model for the marker and the event. Generally, joint models assume that there is a latent process V, and that both marker measurements and the risk of the event depend on V. One main assumption is that conditionally on V, observations of the marker and time of the event are independent. V may be a nondegenerate process such as a Brownian motion with drift or degenerate, depending on only a finite number of random effects such as in shared random-effect models (Rizopoulos, 2010) or JLCMs. A joint model can be specified by a family of densities $(g^{\beta}_{T,Z,V})_{\beta \in B}$, $B \subset \Re^d$, where $\beta = (\alpha, \theta, \xi)$ which can be decomposed in terms of $g^{\alpha}_{T|V}$, $g^{\theta}_{Z|V}$ and g^{ξ}_{V}. The whole information $\bar{\mathcal{O}}_n = (\tilde{T}_i, \delta_i, \bar{Z}^i_{\tilde{T}_i}, i = 1, \ldots, n)$, where $\bar{Z}^i_{\tilde{T}_i}$ is the whole vector of covariates and marker measurements up to $\tilde{T}_i$, is used for estimation. Thus, with the joint model approach, there is only one model fitted with the whole information of the sample, and the whole information of a subject up to t can be used for prognosis at t. That is, we will use $g^{\hat{\beta}_n}_{T|\bar{Z}_t, \tilde{T}>t}$, where $\hat{\beta}_n$ is the maximum likelihood estimator (MLE), or more generally a M-estimator, of β.

2.2 *Expected Cross-Entropy and Kullback–Leibler Risk for Prognostic Estimators*

Conventional quantities in information theory are the entropy, the Kullback–Leibler divergence, and the cross-entropy; see Cover and Thomas (1991). These quantities have been adapted to describe the risk of an estimator of the true density function f^* (Commenges et al., 2008; Liquet and Commenges, 2011). Denoting $g^{\hat{\beta}_n}$ such an estimator, the expected Kullback–Leibler divergence can be defined as: $\mathrm{EKL}(g^{\hat{\beta}_n}) = \mathrm{E}_*[\log \frac{f^*(T)}{g^{\hat{\beta}_n}(T)}]$. Here it is implicit that the expected divergence is with respect to f^* and E_* is short for E_{f^*}; in the following E_* will always mean expectation under the true law. Similarly we define the expected cross-entropy

$\mathrm{ECE}(g^{\hat{\beta}_n}) = \mathrm{E}_*[-\log g^{\hat{\beta}_n}(T)]$. Cross-entropy can be additively decomposed as $\mathrm{ECE}(g^{\hat{\beta}_n}) = \mathrm{EKL}(g^{\hat{\beta}_n}) + \mathrm{H}(f^*)$, with the interpretation that the risk with $g^{\hat{\beta}_n}$ is the risk incurred in using $g^{\hat{\beta}_n}$ in place of f^* plus the risk using f^* . We will adapt these concepts to the prognosis context.

Any prognosis should ideally be made using this (true) prognostic density $f^*_{T|\bar{Z}_t,T>t}$. The accuracy of this ideal prediction depends on the spread of the prognostic density. This could be summarized by the standard deviation of the density. A more general measure, however, is the (relative) entropy of this density. Here the object of interest is a conditional density on both $T > t$ and $\bar{Z}_t$. This conditional density is considered as a random quantity depending on the value of $\bar{Z}_t$ so that we will work with an expected conditional entropy: $\mathrm{EH}^{\bar{Z}_t}(f^*(T|\bar{Z}_t, T > t)) = \mathrm{E}_*[-\log f^*(T|\bar{Z}_t, T > t)|T > t]$. This quantity remains conditional on $\{T > t\}$ because we are not interested in the prognosis for $T < t$.

For prognosis we may use any estimator $g^{\hat{\beta}_n}_{T|\bar{Z}_t,T>t}$ of $f^*_{T|\bar{Z}_t,T>t}$, which may be obtained from a joint model or from another method. We will consider the case where prognosis estimators are chosen as minimizing an estimating function $\Phi_{\bar{\mathcal{O}}_n}(\beta)$, where $\Phi_{\bar{\mathcal{O}}_n}(\beta) = n^{-1}\sum_{i=1}^n \phi_i(\beta)$. In this case, $\hat{\beta}_n$ is a M-estimator that converges under rather weak conditions toward β_0 which minimizes $\mathrm{E}_*[\phi_i(\beta)]$ (Van der Vaart, 2000). Assuming continuity of $g^\beta(u)$ in β we have also that $g^{\hat{\beta}_n}(u)$ converges toward $g^{\beta_0}(u)$ for all u. It is not assumed that the model is well specified: g^{β_0} is the best distribution in the model for risk $\mathrm{E}_*[\phi_i(\beta)]$. The question is then to assess the risk associated to $g^{\hat{\beta}_n}_{T|\bar{Z}_t,T>t}$. In information theory, the loss function is minus the log of the density of the observed variables. The risk is the expectation of the loss function, but here we take into account that the risk is assessed only on $T > t$; thus we take the conditional expectation for defining the EPCE:

$$\mathrm{EPCE}(g^{\hat{\beta}_n}, t) = \mathrm{E}_*\big[-\log g^{\hat{\beta}_n}(T|\bar{Z}_t, T > t)|T > t\big].$$

Note that $\mathrm{EPCE}(g^{\hat{\beta}_n}, t)$ is not a random variable. The expression $\mathrm{E}_*[\log g^{\hat{\beta}_n}(T|\bar{Z}_t, T > t)|T > t]$ is the value of the conditional expectation $\mathrm{E}_*[\log g^{\hat{\beta}_n}(T|\bar{Z}_t, T > t)|1_{T>t})]$ taken on the event $\{T > t\}$. One can highlight here the contrast with Akaike criterion which estimates the global cross-entropy $\mathrm{E}_*[-\log g^{\hat{\beta}_n}_{T,Z}(T, Z)]$ rather than the prognosis cross-entropy.

One can also define the risk from the Kullback–Leibler divergence of the estimator from the true conditional distribution. The expected prognosis Kullback–Leibler risk is

$$\mathrm{EPKL}(g^{\hat{\beta}_n}, t) = \mathrm{E}_*\left[\log \frac{f^*(T|\bar{Z}_t, T > t)}{g^{\hat{\beta}_n}(T|\bar{Z}_t, T > t)}|T > t\right].$$

EPCE is an absolute risk that we can more directly estimate than EPKL but which has not received much interpretation yet. Similarly, as for ECE we have the relation:

$$\mathrm{EPCE}(g^{\hat{\beta}_n}, t) = \mathrm{EPKL}(g^{\hat{\beta}_n}, t) + \mathrm{EH}^{\bar{Z}_t}(f^*(T|\bar{Z}_t, T > t).$$

Thus the absolute risk $\mathrm{EPCE}(g^{\hat{\beta}_n}, t)$ is the sum of the divergence of the prognosis estimator relative to the truth plus the risk using the true prognostic density. Because the divergence is always nonnegative, it follows that EPCE is minimized by f^* and that $\mathrm{EH}^{\bar{Z}_t}(f^*(T|\bar{Z}_t, T > t))$ is the smallest achievable risk. Note that for a continuous variable the entropy is not necessarily positive.

For comparison of prognosis estimators, it is immaterial to use EPCE or EPKL because the difference of risks between two prognosis estimators $g^{\hat{\beta}_n}$ and $h^{\hat{\gamma}_n}$ can be characterized by a unique quantity:

$$\begin{aligned}\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}, t) &= \mathrm{EPCE}(g^{\hat{\beta}_n}, t) - \mathrm{EPCE}(h^{\hat{\gamma}_n}, t)\\ &= \mathrm{EPKL}(g^{\hat{\beta}_n}, t) - \mathrm{EPKL}(h^{\hat{\gamma}_n}, t).\end{aligned}$$

3. Estimation of the Prognostic Risk by Cross-validation: Uncensored Observations

3.1 *Mean Prognosis Loss*

Assume that we wish to use prognosis estimator $g^{\hat{\beta}_n}$, where $\hat{\beta}_n$ is based on a sample of n observations, $\bar{\mathcal{O}}_n = (T_i, Z^i_{T_i}, i = 1, \ldots, n)$, that we assume uncensored in this section. The prognosis estimator may be the MLE in a model $(g^\beta)_{\beta\in B}$. It is not assumed, however, that the model is well specified and the theory may also apply to M-estimators and penalized likelihood estimators. A natural estimator of $\mathrm{EPCE}(g^{\hat{\beta}_n}, t)$ is the mean prognostic loss:

$$\mathrm{MPL}(g^{\hat{\beta}_n}, t) = \frac{1}{N_t}\sum_{i=1}^n G_i(\hat{\beta}_n, t),$$

where $G_i(\beta, t) = -1_{T_i>t}\log(g^\beta(T_i|Z^i_t, T_i > t))$ and $N_t = \sum 1_{T_i>t}$. If $\hat{\beta}_n$ is consistent for β_0, $\mathrm{EPCE}(g^{\hat{\beta}_n}, t)$ converges in probability toward $\mathrm{EPCE}(g^{\beta_0}, t)$ (which could be called simply "prognostic cross-entropy" as g^{β_0} is deterministic). By the law of large numbers and the continuous mapping theorem, $\mathrm{MPL}(g^{\hat{\beta}_n}, t)$ also converges toward $\mathrm{EPCE}(g^{\beta_0}, t)$. However, if n is not very large, it is necessary to take the variability of $\hat{\beta}_n$ into account. $\mathrm{MPL}(g^{\hat{\beta}_n}, t)$ underestimates $\mathrm{EPCE}(g^{\hat{\beta}_n}, t)$ because $g^{\hat{\beta}_n}$ minimizes the log-likelihood. The bias can be studied from equation (3) in Section 3.3.

3.2 *Cross-Validated Prognosis Loss: CVPL*

We consider the leave-one-out LCV criterion as a possible "estimator" of $\mathrm{EPCE}(g^{\hat{\beta}_n}, t)$. It is defined as

$$\mathrm{CVPL}(g^{\hat{\beta}_n}, t) = \frac{1}{N_t}\sum_{i=1}^n G_i(\hat{\beta}_{-i}, t),$$

where $\hat{\beta}_{-i} = \mathrm{argmin}[\Phi_{\bar{\mathcal{O}}_{n|i}}]$ and $\bar{\mathcal{O}}_{n|i}$ is the total observation diminished of observation i.

Then, we define an estimator of the difference of risks of two prognosis estimators $\hat{\gamma}_n$:

$$\mathrm{D}_{\mathrm{CVPL}}(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}, t) = \mathrm{CVPL}(g^{\hat{\beta}_n}, t) - \mathrm{CVPL}(h^{\hat{\gamma}_n}, t), \quad (1)$$

$\mathrm{CVPL}(g^{\hat{\beta}_n}, t)$ has the following property:

$$\mathrm{E}_*[\mathrm{CVPL}(g^{\hat{\beta}_n}, t)] = \mathrm{EPCE}(g^{\hat{\beta}_{n-1}}, t). \quad (2)$$

That is, CVPL estimates without bias the risk that we would have with a sample size of $n-1$. The proof is as follows:

$$\begin{aligned}
&\mathrm{E}_*\big[\mathrm{CVPL}(g^{\hat{\beta}_n},t)|1_{T_1>t},\ldots,1_{T_n>t}\big]\\
&\quad=\mathrm{E}_*\Bigg[\frac{1}{N_t}\sum_{i=1}^{n}-1_{T_i>t}\log\left(g^{\hat{\beta}_{-i}}\left(T_i|Z_t^i,T_i>t\right)\right)\\
&\qquad|1_{T_1>t},\ldots,1_{T_n>t}\Bigg]\\
&\quad=-\mathrm{E}_*[\log g^{\hat{\beta}_{n-1}}(T|Z,T>t)]\\
&\quad=\mathrm{EPCE}(g^{\hat{\beta}_{n-1}},t).
\end{aligned}$$

Because $\mathrm{EPCE}(g^{\hat{\beta}_{n-1}},t)$ does not depend on $1_{T_1>t},\ldots,1_{T_n>t}$, we obtain equation (2). It is intuitive that the difference between the risks with n and $n-1$ is small. Under some regularity conditions it can be proved to be in $O(n^{-3/2})$ (see Supplementary Material).

3.3 *Approximate Formula for* CVPL: CVPL_a

The leave-one-out cross-validation however may be computationally demanding. Thus, it is interesting to develop an approximate formula. This can be done by adapting a result of Commenges et al. (2007), Section 6.2.2. We obtain:

$$\mathrm{CVPL}(g^{\hat{\beta}},t)=\mathrm{MPL}(g^{\hat{\beta}},t)+\mathrm{Trace}\big(H_{\Phi_{\bar{\mathcal{O}}_n}}^{-1}K_t\big)+O_p(n^{-2}), \tag{3}$$

where $H_{\Phi_{\bar{\mathcal{O}}_n}}=\frac{\partial^2\Phi_{\bar{\mathcal{O}}_n}}{\partial\beta^2}$, $K_t=\frac{1}{N_t}\sum_{i=1}^{n}\hat{v}_i(t)\hat{d}_i^T$, with $\hat{d}_i=\frac{1}{n-1}\frac{\partial\phi_i}{\partial\beta}|_{\hat{\beta}_n}$ and $\hat{v}_i(t)=\frac{\partial G_i(\beta,t)}{\partial\beta}|_{\hat{\beta}_n}$. The proof is given in Supplementary material. Note that $\mathrm{CVPL}(g^{\hat{\beta}},t)$ and $\mathrm{MPL}(g^{\hat{\beta}},t)$ differ by the term $\mathrm{Trace}(H_{\Phi_{\bar{\mathcal{O}}_n}}^{-1}K_t)$ which is nonnegative and an $O_p(n^{-1})$. We define CVPL_a as

$$\mathrm{CVPL}_a(g^{\hat{\beta}},t)=\mathrm{MPL}(g^{\hat{\beta}},t)+\mathrm{Trace}\big(H_{\Phi_{\bar{\mathcal{O}}_n}}^{-1}K_t\big). \tag{4}$$

4. Estimation of the Prognostic Expected Observed LCV: Censored Observations

In event history analysis, there are often right-censored observations $(\tilde{T}_i,\delta_i)$. What is often done is to replace the criterion which requires exact observations by its expectation given the censored observations. This is the route followed by Schemper and Henderson (2000), Henderson et al. (2002), and Proust-Lima and Taylor (2009). It would be possible to do it for our score for estimating EPCE. However, for computing the expectation we need to know the distribution of the event. Replacing this distribution by an estimate may be dangerous. If the model is strongly misspecified, the score which assesses its performance may be badly computed, and the approach is likely to produce biased estimators of the risk. Another route taken by Schoop et al. (2008) uses inverse probability weighting. A first estimator uses the Kaplan–Meier estimator of the censoring distribution but is correct only under the strong assumption of a completely independent censoring; a second estimator could use a model for the censoring time taking into account the marker path following Gerds and Schumacher (2006) proposal. However, first, such a model may be misspecified; second, inverse probability weighting estimates may be unstable; third, derived predictive accuracy measures may lack efficiency (Rosthoj and Keiding, 2004).

Thus, we adopt an approach which is in fact commonly used with censored data and which relies on the likelihood of the observation. This was investigated by Commenges et al. (2007) but we have to adapt it to the context of prognosis. Here the loss function will be $-L^{g^{\hat{\beta}_n}}_{\mathcal{O}|\bar{Z}_t,T>t}$, that is, minus the log-likelihood of the observation for $g^{\hat{\beta}_n}$ ignoring the mechanism leading to incomplete data, which is computed as if the censoring times were fixed. This is relevant in the case called CAR(TCMP) by Commenges et al. (2007) (Section 4) which is here identical to "independent censoring" defined in Andersen et al. (1993) (Definition III.2.1). In that case, because of a factorization of the likelihood the term relative to the observation of the censoring can be dropped.

The relevant risk function is then the EPOCE: $\mathrm{EPOCE}(g^{\hat{\beta}_n},t)=\mathrm{E}_*[-L^{\hat{\beta}_n}_{\mathcal{O}|\bar{Z}_t,T>t}|\tilde{T}>t]$. The additive decomposition in divergence plus entropy still holds and $\mathrm{EPOCE}(f^*,t)$ is the smallest achievable risk. $\mathrm{EPOCE}(g^{\hat{\beta}_n},t)$ reduces to $\mathrm{EPCE}(g^{\hat{\beta}_n},t)$ if there is no censoring. This is less easily interpretable in case of censoring but can be readily computed because by definition the likelihood can be computed from the observations. Thus, we do not need to estimate the distribution of the censoring variable. EPOCE, however, still depends on this distribution. Thus it can be considered as a measure of the prognosis risk only for situations with the same censoring distribution. The same formulas as above can be used for defining estimators of EPOCE: mean prognostic observed loss (MPOL) and the cross-validated prognosis observed loss (CVPOL) as well as its approximation CVPOL_a. The formulas are the same, with G_i equal to minus the conditional log-likelihood: $G_i(\beta,t)=-1_{\tilde{T}_i>t}L^{g^{\beta}}_{\mathcal{O}_i|Z_t^i,T_i>t}$ (which is equal to the log of the conditional density if there is no censoring). In this case, $N_t=\sum_{i=1}^{n}1_{\tilde{T}_i>t}$. We can use $D_{\mathrm{CVPOL}}(g^{\hat{\beta}_n},h^{\hat{\gamma}_n},t)$ for estimating $\Delta(g^{\hat{\beta}_n},h^{\hat{\gamma}_n},t)=\mathrm{EPOCE}(g^{\hat{\beta}_n},t)-\mathrm{EPOCE}(h^{\hat{\gamma}_n},t)$.

5. Asymptotic Distribution and Tracking Interval

As in Commenges et al. (2008), it is possible to give the asymptotic distribution of D_{CVPL_a} or D_{CVPOL_a} and deduce a tracking interval from it. The term "tracking interval" is used rather than "confidence interval" because the target $\hat{\gamma}_n$ moves with n . Here, the asymptotics for $D_{\mathrm{CVPOL}_a}(g^{\hat{\beta}_n},h^{\hat{\gamma}_n},t)$ is when N_t (and of course n) tends toward ∞. We focus on the case where $g^{\beta_0}\neq h^{\gamma_0}$. We obtain in that case

$$N_t^{1/2}\big[D_{\mathrm{CVPOL}_a}(g^{\hat{\beta}_n},h^{\hat{\gamma}_n},t)-\Delta(g^{\hat{\beta}_n},h^{\hat{\gamma}_n},t)\big]\xrightarrow{D}\mathcal{N}\big(0,\omega_{t*}^2\big), \tag{5}$$

where $\omega_{t*}^2=\mathrm{var}[G_i(\beta_0,t)-H_i(\gamma_0,t)]$, with $H_i(\gamma,t)=-1_{\tilde{T}_i>t}L^{h^{\gamma}}_{\mathcal{O}_i|Z_t^i,T_i>t}$. The proof is easy because the correction term is asymptotically negligible so that the asymptotic distribution is the same as that of the log of a likelihood ratio MPOL; thus the result applies to differences of MPOL and CVPOL as well as for CVPL_a, MPL, and CVPL in the uncensored case. Beware, however, that the distribution of the likelihood ratio is different in the case $g^{\beta_0}_{T|Z,T>t}=h^{\gamma_0}_{T|Z,T>t}$ (the classical result used for the "likelihood ratio test") and the case

$g^{\beta_0}_{T|Z,T>t} \neq h^{\gamma_0}_{T|Z,T>t}$, as shown by Vuong (1989). A natural estimator of ω^2_{t*} is

$$\hat{\omega}^2_{tn} = N_t^{-1}\sum_{i=1}^{n}[G_i(\hat{\beta}_n,t) - H_i(\hat{\gamma}_n,t)]^2 - \left[N_t^{-1}\sum_{i=1}^{n} G_i(\hat{\beta}_n,t) - H_i(\hat{\gamma}_n,t)\right]^2.$$

From this, we can compute the tracking interval (A_{tn}, B_{tn}), where $A_{tn} = D_{\text{CVPOL}_a}(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}, t) - z_{\alpha/2}N_t^{-1/2}\hat{\omega}_{tn}$ and $B_{tn} = D_{\text{CVPOL}_a}(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}, t) + z_{\alpha/2}N_t^{-1/2}\hat{\omega}_{tn}$, where $1 - \Phi(z_{\alpha/2}) = \alpha/2$ and Φ is the cumulative distribution function of the standard normal variable. The assumption $g^{\beta_0}_{T|Z,T>t} \neq h^{\gamma_0}_{T|Z,T>t}$ is necessarily the case if the models do not overlap and may also be often the case even if the models overlap or are nested. However, in the latter case the convergence toward the normal may be slow and it may be desirable to construct tracking intervals compatible with the likelihood ratio test. This could also be done along the lines of Commenges et al. (2008). Even in this case $\hat{\omega}^2_{tn}$ still gives an idea of the variability of D_{CVPOL_a}.

6. Extensions

6.1 *Extension to the Penalized Likelihood Approach*

The criteria derived for prognosis estimators based on parametric models can be extended to penalized likelihood estimators. For instance, the JLCM model described in Section 7.2 introduces Weibull baseline hazard functions λ_{0v} for latent classes $v = 1, \ldots, K$. Penalized likelihood allows relaxing the Weibull assumption. Assuming the λ_{0v} smooth, a family of estimators is defined as maximizing the penalized log-likelihood: $n^{-1}L_{\bar{\mathcal{O}}_n} - \kappa J(\lambda_{0v}, v = 1\ldots, K)$, where $J(\lambda_{0v}, v = 1\ldots, K)$ is a penalty term taking high value for nonsmooth λ_{0v}'s (for instance the norm of second derivatives). Consistency results have been obtained for $\kappa \to 0$ in a suitable way; see, for instance, Cox and O'Sullivan (1996). Penalized likelihood estimators minimize $\Phi_{\bar{\mathcal{O}}_n}(\beta) = n^{-1}\sum_{i=1}^{n}\phi_i(\beta)$, where $\phi_i(\beta) = -L_{\mathcal{O}_i} + n^{-1}\kappa J(\lambda_{0v}, v = 1\ldots, K)$. The λ_{0v}'s are represented on a basis of splines by a parameter vector γ. Denoting by θ the vector of parameters (ordinary parameters plus spline parameters), we can write the family of estimators as $g^{\hat{\theta}_\kappa}, \kappa > 0$. The risk criteria as well as the cross-validation estimators can be applied to these estimators and can be used, in line with Commenges et al. (2007), for both purposes: choosing the structure of the model (number of classes in a JLCM for instance) and choosing the smoothing parameter κ. The approximate cross-validation criterion (4) can be applied, although the order of the error terms cannot be guaranteed to be the same in this much more complex asymptotics.

6.2 *The Integrated EPCE (IEPCE) and EPOCE and Their Estimators*

We most often wish to make prognosis at different times. When time t of prognosis changes the quantity of information of the marker Z, the distribution of the event to be predicted, as well as the quantity of information for building the prognosis estimator change. It may thus happen that the best prognosis estimators for different times are not based on the same model, which may be considered as undesirable from a practical point of view. A global criterion, called IEPCE, can be constructed by integrating $\text{EPCE}(g^{\hat{\beta}_n}, t)$ on a given distribution of prediction times. This can be approximated by a weighted mean over M time points leading to the criterion: $\text{IEPCE} = \frac{1}{w_.}\sum_{m=1}^{M} w_m \,\text{EPCE}(g^{\hat{\beta}_n}, t_m)$, where $w_. = \sum_{m=1}^{M} w_m$. Estimators of IEPCE and its extension to censored data, IEPOCE, are obtained by the weighted mean applied to equation (4): $\text{ICVPOL}_a = \frac{1}{w_.}\sum_{m=1}^{M} w_m[\text{MPOL}(g^{\hat{\beta}}, t_m) + \text{Trace}(H^{-1}_{\Phi_{\bar{\mathcal{O}}_n}} K_{t_m})]$. Tracking intervals can also be constructed. Neglecting the variability of the trace, we are led to study that of $\frac{1}{w_.}\sum_{m=1}^{M} w_m \,\text{MPOL}(g^{\hat{\beta}}, t_m) = \frac{1}{w_.}\sum_{m=1}^{M}\frac{w_m}{N_{t_m}}\sum_{i=1}^{n} G_i(\hat{\beta}_n, t_m)$ which can be written as $n^{-1}\sum_{i=1}^{n}\text{IG}_\text{i}$, where $\text{IG}_\text{i} = \frac{1}{w_.}\sum_{m=1}^{M}\frac{n w_m}{N_{t_m}} G_i(\hat{\beta}_n, t_m)$. Thus the variance of a difference of IEPOCE can estimated by $n^{-1}\widehat{\text{var}}_*(\text{IG}_\text{i} - \text{IH}_\text{i})$, where $\text{IH}_\text{i} = \frac{1}{w_.}\sum_{m=1}^{M}\frac{n w_m}{N_{t_m}} H_i(\hat{\gamma}_n, t_m)$ and $\widehat{\text{var}}_*(\text{IG}_\text{i} - \text{IH}_\text{i})$ is the empirical variance of $\text{IG}_\text{i} - \text{IH}_\text{i}$. Possible choices for the weights are $w_m = 1$ or $w_m = N_{t_m}$.

7. Illustration: Choice of JLCMs for Prostate Cancer Recurrence

7.1 *The Data Set*

The data set consisted of 459 patients treated by EBRT for localized prostate cancer at the University of Michigan between 1988 and 2004. After the end of EBRT, patients were followed up until clinical recurrence of prostate cancer or last contact with repeated PSA values collected after the end of EBRT and until clinical recurrence. Among the 459 men, 74 men (16.1%) underwent a clinical recurrence during the follow-up with a median time to recurrence of 2.77 years; see Taylor et al. (2005); Jacqmin-Gadda et al. (2010) for details. We aimed at predicting clinical recurrence.

7.2 *The JLCM*

We fitted JLCMs proposed by Proust-Lima and Taylor (2009), as well as standard proportional hazard models that included only prognostic factors available at diagnosis. In brief, a JLCM with K classes assumes that the population consists of a mixture of K homogeneous subpopulations. The latent class membership is defined by the latent variable V_i (see Section 2.1) which takes values $1, \ldots, K$. We assume for simplicity that the marginal probability that $V_i = v$ does not depend on covariates and, with the notation of Section 2.1, is $g^{\xi}(v) = \frac{exp(\xi_v)}{\sum_{l=1}^{K} exp(\xi_l)}$ for $v = 1, \ldots, K$ and $\xi_K = 0$.

For every subject i, measurements of PSA (in ng/ml) were done at times $(t_{i1}, \ldots, t_{ij}, \ldots, t_{in_i})$ previous to cancer recurrence. These measurements were transformed as $Z^{ij} = \ln(\text{PSA}_i(t_{ij}) + 0.1)$ (to get closer to the Gaussian assumption of the linear mixed model below). Previous analyses of this data set (Proust-Lima et al., 2008) showed that post-treatment evolution of $\ln(\text{PSA}(t) + 0.1)$ exhibited a decline in the first years after EBRT, and a subsequent stable or increasing linear trend. So the class-specific linear mixed model for PSA trajectory was defined by:

$$Z^{ij}\,|_{V_i=v} = u_{0iv} + u_{1iv} f_1(t_{ij}) + u_{2iv} t_{ij} + \epsilon_{ij},$$

where $f_1(t) = (1+t)^{-1.5} - 1$; $\epsilon_{ij} \sim_{iid} \mathcal{N}(0, \sigma^2)$; and $u_{iv} = (u_{0iv}, u_{1iv}, u_{2iv})^T \sim \mathcal{N}((\mu_{0iv}, \mu_{1iv}, \mu_{2iv})^T, \omega_v^2 B)$, B being an

Table 1
CVPOL$_a$ *for prognosis estimators based on JLCM's with varying number of latent classes K (from 1 to 5) and different times of prediction t from 1 to 6 years after the end of radiation therapy. The risk set at time t (N_t) and the percentage of censored observation among the risk set are also given*

t	N_t (% censoring)	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
1	459 (83.9%)	0.6143	0.6074	0.5821	0.5794	0.5857
2	386 (86.3%)	0.5682	0.5621	0.5123	0.5172	0.5154
3	327 (89.3%)	0.4787	0.4331	0.3824	0.3823	0.3844
4	285 (93.3%)	0.3504	0.3270	0.3272	0.3284	0.3299
5	237 (92.4%)	0.3678	0.3495	0.3501	0.3520	0.3528
6	184 (92.9%)	0.3410	0.3386	0.3589	0.3617	0.3601

unstructured covariance matrix and $\omega_K = 1$ for identifiability.

The time-to-event model was defined as $\lambda^i(t\,|_{V_i=v}) = \lambda_{0v}(t)\exp(\beta X_i)$, that is a proportional hazard model with class-specific Weibull baseline risk functions λ_{0v} and three pretreatment prognostic factors as fixed covariates with common effects over classes (the X_i's): Gleason score (a scale that measures grades of prostate cancer) in three categories (7 and 8–10 versus 2–6), T-stage category (3–4 versus 1–2), and the pretreatment level of PSA transformed to ln(PSA + 0.1). We chose a relatively simple JLCM to illustrate the method but a more sophisticated JLCM could be evaluated in the same way. Class-specific Weibull risk functions were chosen in a preliminary analysis that highlighted an equivalent fit of the data when assuming Weibull risk functions rather than piecewise constant or M-splines risk functions. Models were estimated using the Jointlcmm function within lcmm R package. With a single latent class ($K = 1$), the JLCM assumes independence between PSA repeated measures and the risk of recurrence so that the survival part of the model reduces to the standard proportional hazard model with prognostic factors available at diagnosis.

In a JLCM, the individual contribution to the conditional log-likelihood required for CVPOL$_a$ computation, with the independent censoring assumption, is:

$$G_i(\beta,t) = -1_{T_i>t}\log\left(\frac{\sum_{v=1}^{K}\left[g^{\alpha}_{T|V,X_i}(\tilde{T}_i|v,X_i)\right]^{\delta_i}\left[\int_0^{\tilde{T}_i} g^{\alpha}_{T|V,X_i}(u|v,X_i)du\right]^{1-\delta_i} g^{\theta}_{Z}(Z_{it}|v)g^{\xi}(v)}{\sum_{v=1}^{K}\left[\int_0^{t} g^{\alpha}_{T|V,X_i}(u|v,X_i)du\right] g^{\theta}_{Z}(Z_{it}|v)g^{\xi}(v)}\right). \qquad (6)$$

Here, $g^{\alpha}_{T|V,X_i}(.|v,X_i)$ is the conditional density function derived from a Weibull proportional hazard model.

7.3 EPOCE *Estimates*

Table 1 provides CVPOL$_a$ for prognosis estimators based on JLCMs with varying number of latent classes (from one to five) from a time of prediction between 1 and 6 years after radiation therapy. We note that the prognosis estimator derived from one latent class model does not make use of longitudinal measures of PSA, and is equivalent to the standard proportional hazard model. The CVPOL$_a$ values are represented

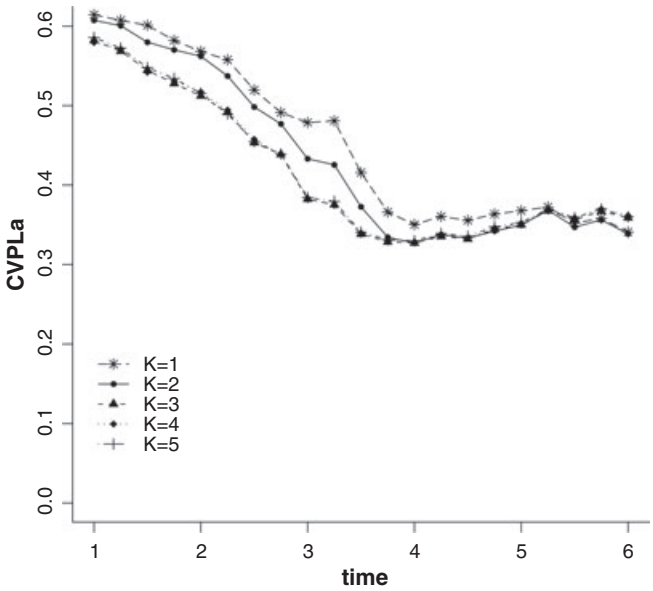

Figure 1. CVPOL$_a$ for prognosis estimators based on JLCM's with varying number of latent classes K (from 1 to 5) and different prognosis times t from 1 to 6 years after the end of radiation therapy.

over time in Figure 1. For any prognosis time, the prognosis estimator based on the JLCM with two latent classes provided a better CVPOL$_a$ than that based on the standard proportional hazard model. Up to almost 4 years after the end of radiation therapy, the models with three or four latent classes gave a smaller CVPOL$_a$ than simpler models with one or two latent classes, indicating a better prognosis accuracy. This was confirmed by the differences in CVPOL$_a$ with 95% tracking intervals plotted in Figure 2. For predictions from 1 to 3 years after radiation therapy, the model with three latent classes gave a better CVPOL$_a$ with tracking interval excluding 0 than the model with two latent classes. In contrast, from 3.5 years after radiation therapy, some point estimates of prognostic performances of the model with two latent classes were better than those for the model with three latent classes. However, when looking at the tracking interval it cannot be concluded which prognosis estimator is the best and the differences are anyhow small. The prognosis estimators based on models with four or five latent classes show no significant improvement. Thus, from a practical point of view it is reasonable to recommend the prognosis estimator based on the three-class model for all prognosis times. This conclusion can be reached directly by computing the criterion ICVPOL$_a$ which estimates the integrated criterion IEPOCE. With $M = 21$

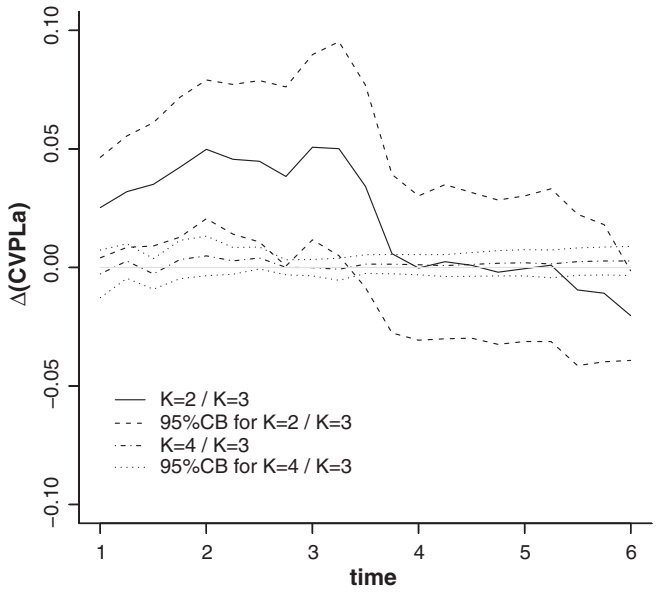

Figure 2. Differences in CVPOL_a with 95% tracking interval at different prognosis times (from 1 to 6 years after the end of radiation therapy) for prognosis estimators based on JLCM's with $K = 2$ compared to $K = 3$ latent classes or $K = 3$ compared to $K = 4$ latent classes.

and $w_m = 1$, we found $0.453, 0.433, 0.414, 0.415,$ and 0.416 for $K = 1, 2, 3, 4,$ and 5, respectively. Thus, according to the integrated criterion, $K = 3$ leads to the best prognosis estimator. The tracking interval for the differences of IEPOCE for $K = 3$ and its two neighbors are $[-0.004; 0.002]$ and $[-0.044; 0.004]$ for $K = 4$ and $K = 2$, respectively. We conclude that there cannot be much difference between $K = 3$ and $K = 4$; we are not sure that the prognosis estimator based on $K = 3$ is better than that based on $K = 2$ (as zero is in the tracking interval). However, there could be a difference of order 0.04 in favor of $K = 3$.

8. Conclusion

Our information-based criteria are easy to compute, and have good properties without making strong assumptions on the censoring mechanism. These properties have been studied by simulation for both well-specified and misspecified models, and found to be satisfactory; see Supplementary Material. The criteria allow assessing estimators of the whole distribution of the time of event. For simpler communication with clinicians, it may be interesting to predict whether the event will occur before a certain time $t + s$. The theory also applies replacing $f^*(T|\bar{Z}_t, T > t)$ by $f^*(Y|\bar{Z}_t, T > t)$, where $Y = 1_{t<T\leq s}$. In the latter case, if we consider different values of s we may apply the method repeatedly for the different values. Another approach is to choose the best estimator for the continuous variable, and then deduce from this estimator the probability that the event occurs before s. This approach uses more information but often requires more parameters and assumptions. Liquet and Commenges (2011) compared regression models based on a continuous variable or a dichotomized variable for predicting the dichotomized variable and did not find large differences. So, for making prognosis for windows defined by several values of s, it seems a simpler approach to choose the best estimator for the continuous event time and to deduce the probabilities of the event from it.

The proposed criteria and estimators have been studied in simulation using JLCM's but they can be used for any joint model. Simulations presented in Supplementary material show that the risk estimator CVPOL_a also works when the censoring is not complectly independent but may depend on previous observations.

9. Supplementary Materials

Web Appendices are available under the Paper Information link at the Biometrics website `http://www.tibs.org/biometrics`. They contain the proofs of results given in Section 3 and a simulation study.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, B.N. Petrov and F., Csàri(eds), 267–281. Budapest: Akademiai Kiàdo

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Commenges, D., Joly, P., Gegout-Petit, A., and Liquet, B. (2007). Choice between semi-parametric estimators of Markov and non-Markov multi-state models from generally coarsened observations. *Scandinavian Journal of Statistics* **34,** 33–52.

Commenges, D., Sayyareh, A., Letenneur, L., Guedj, J., and Bar-hen, A. (2008). Estimating a difference of Kullback–Leibler risks using a normalized difference of AIC. *Annals of Applied Statistics* **2,** 1123–1142.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*, page 542. New York: John Wiley and Sons.

Cox, D. and O'Sullivan, F. (1996). Penalized likelihood-type estimators for generalized nonparametric regression. *Journal of Multivariate Analysis* **56,** 185–206.

Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48,** 1029–1040.

Gerds, T. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63**, 1283–1287.

Heagerty, P. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61,** 92–105.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1,** 465–480.

Henderson, R., Diggle, P., and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics* **3**, 33–50.

Jacqmin-Gadda, H., Proust-Lima, C., Taylor, J., and Commenges, D. (2010). Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics* **66,** 11–19.

Lin, H., Turnbull, B., McCulloch, C., and Slate, E. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data. *Journal of the American Statistical Association* **97,** 53–65.

Liquet, B. and Commenges, D. (2011). Choice of estimators based on different observations: Modified AIC and LCV criteria. *Scandinavian Journal of Statistics* **38,** 268–287.

Proust-Lima, C. and Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using

repeated measures of posttreatment PSA: A joint modeling approach. *Biostatistics* **10,** 535–549.

Proust-Lima, C., Taylor, J., Williams, S., Ankerst, D., Liu, N., Kestin, L., Bae, K., and Sandler, H. (2008). Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. *International Journal of Radiation Oncology • Biology • Physics* **72,** 782–791.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35,** 1–33.

Rosthoj, S. and Keiding, N. (2004). Explained variation and predictive accuracy in general parametric statistical models: The role of model misspecification. *Lifetime Data Analysis* **10,** 461–472.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56,** 249–255.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics* **64,** 603–610.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6,** 461–464.

Taylor, J. M. G., Yu, M., and Sandler, H. M. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* **23,** 816–825.

Van der Vaart, A. (2000). *Asymptotic Statistics*. Number 3. Cambridge, UK: Cambridge University Press.

Van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34,** 70–85. doi:10.1111/j.1467-9469.2006.00529.x.

Vuong, Q. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica: Journal of the Econometric Society* **57,** 307–333.

Received March 2011. *Revised September* 2011.
Accepted September 2011.