0646

VALIDATION OF PROGNOSTIC INDICES USING THE FRAITLY MODEL

LEGRAND C., DUCHATEAU L., JANSSENS P., DUCROCQ V., and R. SYLVESTER

Validation of prognostic indices using the frailty model

C. Legrand[1], L. Duchateau[2], P. Janssen[3], V. Ducrocq[4], R. Sylvester[1]

[1] European Organisation for Research and Treatment of Cancer, Av. E. Mounier 83/11, 1200 Brussels, Belgium.
[2] Department of Physiology and Biometrics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, B-9820 Merelbeke, Belgium,
[3] Center for Statistics, Hasselt University, Agoralaan, B-3590 Diepenbeek, Belgium,
[4] Institut National de la Recherche Agronomique, Station de Génétique Quantitative et Appliquée, F-78352 Jouy-en-Josas, France.

Running title: PI validation using frailty models.

Keywords: prognostic index, validation, frailty model, multicenter clinical trial, bladder cancer.

Acknowledgments: This publication was supported by grant number 5U10CA11488-35 and 5U10CA11488-36 from the National Cancer Institute (Bethesda, Maryland USA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute. Paul Janssen acknowledges partial support from the Belgian Federal Science Policy Office (Interuniversity Attraction Pole Programme P5/24).

The authors thank the European Organisation for Research and Treatment of Cancer (EORTC) Genito-Urinary Group for permission to use the data from EORTC trials 30781, 30782, 30791, 30831, 30832, 30845, 30863 for this research.

[1] Corresponding author. Currently at Merck Sharp & Dohme, Inc. Clos du Lynx 5, B-1200 Brussels, Belgium. catherine_legrand@merck.com, Phone: +32 2 7766438, Fax: +32 2 7766186.

Abstract

A prognostic factor analysis is performed to derive, for a particular patient population and for a particular disease, a prognostic index which is used to predict a patient's prognosis and thus aid in determining the most appropriate treatment strategy. A major issue when proposing a new prognostic index is its generalisibility to daily clinical practice. It is therefore recognised that validation is required to assess the generalisibility of a new prognostic index. Validation often consists in assessing how well the prognostic index performs in a new sample of patients (validation sample). However common validation techniques usually only consider whether "on the average" the results obtained by the prognostic index in classifying the patients are the same in the construction set and the validation set. We introduce a new important aspect of the generalisibility of the prognostic index in this paper, namely the homogeneity of the prognostic index risk groups hazard ratios over different centers. If the variability between centers is substantial, the prognostic index may have no discrimination capability in a proportion of the centers. To model such heterogeneity we use a Cox proportional hazards model that includes a random center effect and a random prognostic index by center interaction. Statistical inference for this frailty model is based on a Bayesian approach using a Laplacian approximation for the marginal posterior distribution of the variances of the random effects. Particular attention is drawn to summarizing the information available from this marginal posterior distribution. Our approach is applied to a commonly used prognostic index for bladder cancer patients.

1 Introduction

Prognostic factor models investigate the relationship between patient characteristics and the outcome of the patient. These are typically regression models and the variables found to be associated with the outcome of patients are called prognostic factors. One of the common objectives when identifying such prognostic factors is to determine a prognostic index or score, i.e., a new variable combining information from the identified prognostic factors and whose value can be used to classify patients in different risk groups according to the probability of the event of interest. Such a prognostic index usually takes the form of a weighted sum of the prognostic factors, the weights being derived from the regression coefficients of the model including the identified prognostic factors. For a specific disease, an appropriate prognostic index can then be used by clinicians to adapt the treatment strategy of future patients, e.g., offering a more aggressive treatment to patients at high risk of recurrence.

No widely accepted methodology exists for the design, analysis and interpretation of prognostic factor studies. Numerous issues may cause poor performance of a prognostic index in a new sample of patients (Simon and Altman, 1994; Altman and Royston, 2000). Apart from methodological flaws, it is important to keep in mind that in day-to-day practice a particular prognostic index is rarely used in exactly the same setting as the one in which it has been developed. Therefore, the generalisability of the prognostic index needs to be assessed, a process usually refered to as validation (see Legrand (2005) for an overview of validation techniques). The validation of a prognostic index is a multi-step procedure: the more numerous and diverse the settings in which the prognostic index is shown to perform well, the more the prognostic index is considered to be generalisible to daily clinical practice.

Calibration and discrimination are often considered as the two main components of validation (Justice et al., 1999). Discrimination refers to a prognostic index ability to distinguish patients with different risk. When assessing the discriminatory power of a prognostic index, one merely considers if the relative ranking of individual risk is in the correct order, e.g. observed event rates are higher in patients with higher scores. For time-to-event outcomes, a natural and popular way of investigating the discriminatory power of a prognostic index consists of dividing patients into several risk groups and to look at the graphical display of observed survival curves in the validation sample for the risk groups. Several measures of discrimination have been proposed in the literature (Harrell et al., 1996, Graf et al., 1999; Schemper and Stare, 1996; Schemper and Henderson, 2000; Schemper, 2003). Calibration pertains to the agreement between predicted outcomes and the observed outcomes in the validation sample (van Houwelingen, 2000). Investigating calibration of a new prognostic index thus considers if predicted probabilities are neither too high nor too low.

Besides these important issues, we propose to study an unexplored aspect of prognostic indices, namely its homogeneity over different centers. As noticed by Justice et al (1999), although many prognostic indices are now developed and validated using data from multicenter clinical trials, investigators rarely report the variation in results by center. However, if the results of a prognostic index cannot be reproduced by different clinicians in different centers, its generalisability must be questionned. For a prognostic index which, for a specific time-to-event endpoint, separates patients into two risk groups (e.g., "poor" and "good" prognosis groups), we argue that the heterogeneity of the hazard ratio of the risk groups defined by this prognostic index among different centers provides useful information on its generalisibility. Even if the hazard ratio between prognosis groups associated with a particular prognostic index is similar in the construction and validation set, a large variation in this hazard ratio over centers should be a warning against the generalisibility of the prognostic index. So rather than considering the overall discriminatory power of a prognostic index we are interested in the generalisability of the prognostic index over centers and our objective is to develop tools to quantify this heterogeneity.

To investigate this heterogeneity we use a proportional hazards model that includes the prognostic index as a fixed effect, a random center effect and a random prognostic index by center interaction. We consider a Bayesian approach to fit such a frailty model with two random effects; it is based on the maximisation of the marginal posterior distribution of the variance components, obtained by Laplace approximation. While this approach has already been presented elsewhere (Legrand et al., 2005), we extend this discussion to further investigate how to summarize the information available from the marginal posterior distribution. The proposed method is computationally fast and simulations show that it provides satisfying results when investigating the heterogeneity of a prognostic index, typically considering unequal balance of patients over treatment groups. These new simulations therefore complete the ones previously presented in the context of treatment effect heterogeneity in cancer clinical trials (Legrand et al., 2005) considering then equal distribution of patients over treatment groups.

Section 2 gives the notation and a precise description of the statistical model. In Section 3, we shortly summarize the estimation techniques and propose methods to further explore the information available from the marginal posterior distribution. In Section 4, we illustrate on a real bladder cancer database how investigating heterogeneity in prognostic index effect over centers brings further information regarding generalisability of the prognostic index. Section 5 presents simulations evaluating the performance of our method. Results and findings are discussed in Section 6.

2 Statistical model

Assume that we have data from a total of n patients coming from G different centers, n_i patients coming from center i ($n = \sum_{i=1}^{G} n_i$). For the j^{th} patient in center i, we observe $T_{ij} = \min(Y_{ij}, C_{ij})$ where Y_{ij} is the real time-to-event for this patient and C_{ij} is a random censoring time independent of Y_{ij}. Additionally, a censoring indicator δ_{ij} is observed, δ_{ij} equals 1 if $T_{ij} = Y_{ij}$ and 0 if $T_{ij} = C_{ij}$. For each patient the binary variable x_{ij} indicates whether the patient is classified in the good or poor prognosis group based on the particular prognostic index considered.

Although a wide variety of models are available for the analysis of censored failure time data, the Cox proportional hazards regression model (Cox, 1972) has emerged as the most popular one. We will therefore consider an extension of this model to investigate heterogeneity in the prognostic index risk groups hazard ratio over centers by including, in addition to the prognostic index (PI) indicator x_{ij}, a random center effect b_{0i} and a random effect for the center by PI indicator interaction b_{1i}.

For the j^{th} patient in the i^{th} center we model the hazard as

$$\lambda_{ij}(t \mid \beta, \mathbf{b}) = \lambda_0(t) \exp\Big(b_{0i} + (\beta + b_{1i})x_{ij}\Big) \qquad (1)$$

where $\lambda_0(t)$ represents the unspecified baseline hazard at time t, β is the fixed effect coefficient corresponding to the PI indicator x_{ij}. The factor $\exp(b_{0i})$, with b_{0i} the center effect, represents the deviation of the i^{th} center from the overall underlying baseline risk and its predicted value will therefore be refered to in the following as "the predicted center baseline risk". Similarly, the value of $\exp(b_{1i})$ (interaction term) represents the deviation of the i^{th} center from the overall PI effect $\exp(\beta)$, and we will refer to the predicted value of $\exp(\beta + b_{1i})$ as the "predicted prognostic index effect".

Parallel with mixed models, we assume in (1) that

$$b_{0i} \sim_{iid} N(0, \sigma_0^2)$$

$$b_{1i} \sim_{iid} N(0, \sigma_1^2)$$

with $\{b_{0i}\}$ and $\{b_{1i}\}$ independent.

The variance components of the random effects σ_0^2, and σ_1^2 can be interpreted as a measure of the amount of variation in baseline risk and PI effect over centers respectively.

3 Model fitting

Model (1) has been used to investigate heterogeneity in treatment effect either in the context of multicenter clinical trials (Yamaguchi and Ohashi, 1999; Matsuyama et al., 1998; Yamaguchi et al., 2002; Glidden and Vittinghoff, 2004; Legrand et al., 2005) or in the context of individual patient data meta-analyses (Smith et al., 2005), considering in both cases a random center effect, a fixed treatment effect and a random treatment by center interaction. This model was then fit based on a penalized partial likelihood approach (Yamaguchi and Ohashi, 1999; Smith et al., 2005). However one major drawback of this approach is the long computer-time required to fit this model (Smith et al., 2005). Other estimation methods have been proposed based on the EM algorithm (Ripatti et al., 2002) or on the MCMC method (Vaida and Xu, 2000). None of these approaches has been implemented in a widely available statistical software and the programs made available by the authors are usually very slow.

In a Bayesian context, the inference on the parameters of interest σ_0^2 and σ_1^2 is based on the bivariate marginal posterior distribution obtained after integrating out the fixed and random components from the joint posterior distribution. The marginal posterior distribution can be approximated by Laplacian integration (Ducrocq and Casella, 1996; Legrand et al., 2005). This approach can be summarized as follow, Legrand et al. (2005) provides a more detailed description.

Denoting $\theta^T = (\sigma_0^2, \sigma_1^2)$, the joint posterior density for model (1) is proportional to

$$\pi(\beta, \mathbf{b}, \theta \mid \mathbf{y}) \propto L(\beta, \mathbf{b} \mid \mathbf{y}) \times \pi_0(\mathbf{b} \mid \theta) \times \pi_0(\beta) \times \pi_0(\theta) . \quad (2)$$

where the first factor is the likelihood given the observations, the second factor is the joint prior distribution of the random effects and the last two factors, $\pi_0(\beta)$ and $\pi_0(\theta)$, are the prior distributions of the fixed effect and of the vector of variance components $\theta^T = (\sigma_0^2, \sigma_1^2)$.

Considering a Cox model leaves the baseline hazard unspecified and based on the justification provided by Ibrahim et al. (2001) and Sinha et al. (2003), we use for $L(\beta, \mathbf{b} \mid \mathbf{y})$ the partial likelihood

$$L(\beta, \mathbf{b} \mid \mathbf{y}) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} \left[\frac{\exp\left(b_{0i} + (\beta + b_{1i})\, x_{ij}\right)}{\sum\limits_{t_{kl} \geq t_{ij}} \exp\left(b_{0k} + (\beta + b_{1k})\, x_{kl}\right)} \right]^{\delta_{ij}}.$$

The joint prior distribution of the random effects, assumed to be normal with mean vector $\mathbf{0}$ and variance vector θ (hyperparameter), is given by

$$\pi_0(\mathbf{b} \mid \theta) = \prod_{i=1}^{G} \frac{1}{2\pi\sigma_0\sigma_1} \exp\left(-\frac{1}{2} \left(\frac{b_{0i}^2}{\sigma_0^2} + \frac{b_{1i}^2}{\sigma_1^2} \right) \right). \tag{3}$$

Considering a flat prior distribution for β and θ

$$\pi_0(\theta) \propto 1 \quad \text{and} \quad \pi_0(\beta) \propto 1$$

the log joint posterior density can be written as

$$\begin{aligned} &\log \pi(\beta, \mathbf{b}, \theta \mid \mathbf{y}) \propto \\ &\sum_{i=1}^{G} \sum_{j=1}^{n_i} \delta_{ij} \left[b_{0i} + (\beta + b_{1i})\, x_{ij} - \log \left(\sum_{t_{kl} \geq t_{ij}} \exp\left(b_{0k} + (\beta + b_{1k})\, x_{kl}\right) \right) \right] \\ &-G \log\left(2\pi\sigma_0\sigma_1\right) - \frac{1}{2} \sum_{i=1}^{G} \left(\frac{b_{0i}^2}{\sigma_0^2} + \frac{b_{1i}^2}{\sigma_1^2} \right). \end{aligned} \tag{4}$$

The marginal posterior density of θ is obtained by integrating out the nuisance parameters β and $\mathbf{b}$ from the joint posterior density

$$\pi(\theta \mid \mathbf{y}) = \int \int \pi(\beta, \mathbf{b}, \theta \mid \mathbf{y})\, d\beta d\mathbf{b} \tag{5}$$

where the integration is over R^{2G} for $\mathbf{b}$ and R for β. Laplacian integration (Tierney and Kadane, 1986) can be used to approximate this integral for a particular value θ^* of θ:

$$\pi\left(\theta^* \mid \mathbf{y}\right) \approx (2\pi)^{G+\frac{1}{2}} \left|\hat{\mathbf{H}}_{\theta^*}^{-1}\right|^{\frac{1}{2}} \pi\left(\hat{\mathbf{\Psi}}_{\theta^*} \mid \mathbf{y}, \theta^*\right) \tag{6}$$

where $\mathbf{\Psi} = \left(\beta, \mathbf{b}^T\right)^T$ and for a fixed θ^*, $\hat{\mathbf{\Psi}}_{\theta^*}$ is the mode of the joint posterior for β and $\mathbf{b}$, i.e.,

$$\hat{\mathbf{\Psi}}_{\theta^*} = \left(\hat{\beta}_{\theta^*}, \hat{\mathbf{b}}_{\theta^*}^T\right)^T = \text{Arg}_{\mathbf{\Psi}} \max \pi\left(\mathbf{\Psi} \mid \mathbf{y}, \theta^*\right) \tag{7}$$

and

$$\hat{\mathbf{H}}_{\theta^*} = -\left.\frac{\partial^2 \log \pi\left(\mathbf{\Psi} \mid \mathbf{y}, \theta^*\right)}{\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T}\right|_{\mathbf{\Psi}=\hat{\mathbf{\Psi}}_{\theta^*}}$$

with $\pi(\mathbf{\Psi} \mid \mathbf{y}, \theta^*) = \pi(\beta, \mathbf{b}, \theta^* \mid \mathbf{y})$ (see (2)).

Taking the logarithm on both sides, we can write

$$\begin{aligned} \log \pi\left(\theta^* \mid \mathbf{y}\right) &\approx \text{constant} + \log \pi\left(\hat{\mathbf{\Psi}}_{\theta^*} \mid \mathbf{y}, \theta^*\right) - 0.5 \log \mid \hat{\mathbf{H}}_{\theta^*} \mid \\ &= \text{constant} + f(\theta^*). \end{aligned} \tag{8}$$

In the two-dimensional space of the two variance components, we use the Simplex algorithm (Nelder and Mead, 1965) with σ_0^2 and σ_1^2 as parameters and the approximated marginal posterior log-density (6) as function to identify the values which maximize this approximated marginal posterior distribution. Once these values are found, they are used as estimates of the variance components $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ of the two random effects.

The marginal posterior density contains all the information about the parameters of interest, which we denote $\theta^T = (\sigma_0^2, \sigma_1^2) = (\theta_0, \theta_1)$ for convenience. Besides the mode, it is interesting to get more information on the behavior of this marginal posterior density. In the Bayesian context, the construction of credible sets is based on the whole marginal posterior density of the parameters. In this paper, we are therefore particularily interested in obtaining more information on this density whithout increasing unreasonnably the computation time.

For a grid of points $\theta_{rs} = (\theta_{0r}, \theta_{1s})$, $r = 1, ..., n_r$, $s = 1, ..., n_s$ we can evaluate the value of the approximated marginal posterior density $\pi(\theta_{rs} \mid \mathbf{y})$ according to (6). Considering a sufficiently large number of points, plotting these values provides, after standardisation, a graphical display of the marginal posterior density of $\theta^T = (\theta_0, \theta_1)$. Standardisation is obtained by computing

$$p(\theta_{0r}, \theta_{1s}) = \frac{\pi(\theta_{rs} \mid \mathbf{y})}{\sum_{l=1}^{n_r} \sum_{q=1}^{n_s} \pi(\theta_{lq} \mid \mathbf{y}) \epsilon^2}$$

where $\epsilon = \theta_{0,l+1} - \theta_{0,l} = \theta_{1,q+1} - \theta_{1,q}$ is the distance between two adjacent points, taken to be equidistant.

The values on this grid also allow us to obtain information on the shape of the marginal density of each variance component. Such a "univariate" marginal posterior density is probably more informative for clinicians than the bivariate marginal density of both variance components and allows us to compute credible set for each variance component. To obtain an approximation for the density of θ_0, we can plot the points

$$\tilde{\pi}(\theta_{0r} \mid \mathbf{y}) = \frac{\sum_{s=1}^{n_s} \pi(\theta_{rs} \mid \mathbf{y})}{\sum_{r=1}^{n_r} \sum_{s=1}^{n_s} \pi(\theta_{rs} \mid \mathbf{y}) \epsilon}.$$

For θ_1 we plot $\tilde{\pi}(\theta_{1s} \mid \mathbf{y})$.

However, for databases of the size typically encountered in cancer clinical trials, obtaining the whole marginal posterior density for each variance component using the previous techniques quickly becomes computationaly demanding. Another possibility to obtain an approximation of the marginal posterior density for each of the variance components is to use a Gram-Charlier expansion. Based on the first k moments, a Gram-Charlier expansion of order k approximates the probability density function of a continuous random variable X as an orthogonal expansion derived from the normal distribution.

For any random variable X having a continuous distribution with mean μ and variance σ^2, the density function of the standardised variable $Z = (X - \mu)/\sigma$ can be expanded as (Cramer, 1971)

$$f_Z(z) = c_0\phi(z) + \frac{c_1}{1!}\phi'(z) + \frac{c_2}{2!}\phi''(z) + \ldots \tag{9}$$

where $\phi(.)$ represents the standardised normal density and the c_i are constant coefficients. Furthermore it can be shown that $\phi^{(r)}(z) = (-1)^r H_r(z)\phi(z)$ where $H_r(z)$ is the Hermite polynomial of degree r. Using this result and properties of Hermite polynomials (see Appendix), (9) can be rewritten as

$$\begin{aligned} f_Z(z) &= c_0 H_0(z)\phi(z) - c_1 H_1(z)\phi(z) + \frac{c_2}{2!}H_2(z)\phi(z) - \frac{c_3}{3!}H_3(z)\phi(z) + \ldots \\ &= \phi(z) + \frac{\gamma}{6}(z^3 - 3z)\phi(z) + \ldots \end{aligned}$$

with γ the skewness of Z.

Considering only the first terms leads to the following approximation of the density function of the standardised variable Z

$$f_Z(z) \approx \phi(z)(1 + \frac{1}{6}(z^3 - 3z))$$

Now backtransform to the non-standardised variable X to obtain

$$f_X(x) = \frac{1}{\sigma} f_Z\big(\frac{x-\mu}{\sigma}\big) = \frac{1}{\sigma} f_Z(z). \tag{10}$$

We now use this idea to obtain approximations for the univariate marginal posterior density of each of the variance components.

Denoting $g(\theta_0, \theta_1) = g(\theta) = \exp(f(\theta) - f(\hat{\theta}))$, with $f(.)$ given by (8), we propose to compute the first three mixed moments of each variance component, given by, say for θ_0:

$$\mu_{1:\theta_0} = k \int_0^\infty \int_0^\infty \theta_0 g(\theta_0, \theta_1) d\theta_0 d\theta_1 \tag{11}$$

$$Var_{\theta_0} = \mu_{2:\theta_0} - \mu_{1:\theta_0}^2 = k \int_0^\infty \int_0^\infty \theta_0^2 g(\theta_0, \theta_1) d\theta_0 d\theta_1 - (\mu_{\theta_0})^2 \tag{12}$$

$$\gamma_{\theta_0} = \frac{\mu_{3:\theta_0} - 3\mu_{1:\theta_0}\mu_{2:\theta_0} + 2\mu_{1:\theta_0}^3}{Var_{\theta_0}^{3/2}} \tag{13}$$

$$\text{with} \qquad \mu_{3:\theta_0} = k \int_0^\infty \int_0^\infty \theta_0^3 g(\theta_0, \theta_1) d\theta_0 d\theta_1$$

where k is the appropriate normalizing constant.

Two-dimensional Gaussian quadrature can be used to compute these integrals. After a change of variable to integrate over the adequate range of values $(-\infty, \infty)$, the iterative strategy proposed by Smith et al. (1985) is implemented. After few iterations, this procedure improves the accuracy of the numerical integration and therefore leads to a precise numerical procedure to calculate these moments. Further technical details can be found in Legrand et al. (2005). Finally, we use these moments to calculate the skewness needed in the formula for the Gram-Charlier approximation of the univariate mariginal posterior density of each variance component.

Also, the grid evaluation can be used to obtain a good approximation of the first moments of θ_0 and θ_1 by considering discretised version of (11)-(13), e.g., for θ_0,

$$\begin{aligned} \tilde{\mu}_{\theta_0} &= \sum_{r=1}^{n_r} \theta_{0r}\tilde{\pi}(\theta_{0r} \mid \mathbf{y})\epsilon && (14) \\ &= \frac{\sum_{r=1}^{n_r}\sum_{s=1}^{n_s} \theta_{0r}\pi(\theta_{rs} \mid \mathbf{y})}{\sum_{r=1}^{n_r}\sum_{s=1}^{n_s} \pi(\theta_{rs} \mid \mathbf{y})} && (15) \end{aligned}$$

$$\tilde{V}_{\theta_0} = \tilde{\mu}_{2:\theta_0} - \tilde{\mu}_{\theta_0}^2 = \sum_{r=1}^{n_r} \theta_{0r}^2\tilde{\pi}(\theta_{0r} \mid \mathbf{y})\epsilon - \tilde{\mu}_{\theta_0}^2 \qquad (16)$$

$$\tilde{\gamma}_{\theta_0} = \frac{\tilde{\mu}_{3:\theta_0} - 3\tilde{\mu}_{\theta_0}\tilde{\mu}_{2:\theta_0} + 2\tilde{\mu}_{\theta_0}^3}{(\tilde{V}_{\theta_0})^{3/2}} \qquad (17)$$

with

$$\tilde{\mu}_{3:\theta_0} = \sum_{r=1}^{n_r} \theta_{0r}^3\tilde{\pi}(\theta_{0r} \mid \mathbf{y})\epsilon.$$

We implemented this approach in The Survival Kit V3.12 (Ducrocq and Sölkner, 1994 and 1998), extending the version freely available from the internet[2] to permit the joint estimation of the two variance components and to compute the first three moments of the approximated posterior marginal density of each variance component given by (11)-(13) with an acceptable increase in computation time.

[2] http//www.boku.ac.at/nuwi/software/sofskit.htm

4 Bladder cancer database

Bladder cancer is a common urological malignancy and about 70-80% of all bladder cancers are superficial (stage Ta-T1). We consider a pooled database of seven trials conducted in this patient population by the Genito-Urinary Group of the European Organisation for Research and Treatment of Cancer (EORTC trials 30781, 30782, 30791, 30831, 30832, 30845 and 30863) (Kurth et al., 1984; Bouffioux et al., 1992; Oosterlinck et al., 1993; Bouffioux et al., 1995; Newling et al., 1995; Witjes et al., 1998). These trials were designed to investigate the use of prophylactic treatment following transurethral resection (TUR). All patients randomized had Ta-T1 bladder cancer, approximately half with primary bladder cancer and half with recurrent disease. A total of 2649 eligible patients were included in these trials. However our analysis is restricted to the 2501 patients without missing information for the prognostic index we consider. These patients were recruited by 63 centers.

Prognostic factors in superficial bladder cancer have been the subject of numerous publications over the past years (Sylvester et al., 2006), with the objective of adapting the treatment acording to the risk of the event of interest. In 1998, Allard et al. (1998) developed a prognostic index for disease free interval (DFI) based on a cohort of 382 patients with primary Ta and T1 bladder cancer, of whom 19% of patients received intravesical chemotherapy or immunotherapy during the follow up. Allard et al. (1998) considered the following adverse tumors characteristics (ATCs) present at initial resection: tumor multiplicity, tumor diameter >3 cm, stage T1 and histological grade 2 or 3. They proposed then grouping the patients into four risk groups, each category being simply defined by the number of ATCs: no ATC, 1 ATC, 2 ATCs, 3-4 ATCS.

Our data are not fully comparable in terms of baseline characteristics to the data used in Allard et al. (1998). The Allard prognostic index was originally developed for patients with primary bladder cancer while our database contains 45.2% of recurrent bladder cancer. Furthermore our data also present a much lower proportion of patients with tumors larger than 3 cm (17.8% in our data versus 38.7% in the Allard data). Results in terms of DFI obtained in Allard et al. (1998) and results obtained with the same grouping of patients in our cohort are presented in Table 1. Figure 1 displays Kaplan-Meier estimates of DFI per prognosis group, as obtained in our cohort. Despite a clear loss of calibration (higher percent disease free at 1 and 2 years in most prognostic groups in EORTC data), the prognostic index developed by Allard et al. (1998) still shows good discrimination between prognostic groups when applied to our dataset (independent validation sample).

At this stage of our work, we consider only two risk groups, grouping patients without any ATC at initial resection as good prognosis patients and patients with at least one ATC as poor prognosis patients. This leads to about a 15%-85% distribution of patients over prognostic groups, assigning patients to one of these two risk groups could therefore be used to save one sixth of the patients from more aggressive, more toxic and more expensive treatment. The results in each of these two groups are presented in Table 2. The hazard ratio is 2.09 with a 95% confidence interval well above 1. Fitting model (1) assuming a normal distribution for the center and prognostic index by center random effects, leads to estimates $\hat{\sigma}_0^2 = 0.0953$ and $\hat{\sigma}_1^2 = 0.01619$.

Considering a grid of 200 values for θ_{0r} (from 0 to 0.20 with equidistant span) and 200 values for θ_{1s} (from 0 to 0.200 with equidistant span), we obtain the approximated marginal posterior density shown in Figure 2. The approximations of the univariate marginal density of θ_0 and θ_1 obtained either from the grid evaluation or from univariate Gram-Charlier approximations based on the first three mixed moments are displayed in Figure 3. For each variance component the two approximations appear to be very close. We also computed the values of the first three mixed moments (11)-(13) either based on the grid evaluation or using Gauss-Hermite quadrature (Table 3). Results are very close to the values obtained by numerical integration, except for skewness most probably due to the fact that the estimates of the heterogeneity parameters are close to the boundary of the parameter space

A nice way to interpret the values of $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ is to evaluate their impact on clinically relevant quantities (Legrand et al., 2005; Duchateau and Janssen, 2005). Impact of the value of $\hat{\sigma}_1^2$ on the distribution of hazard ratios $HR = \exp(\beta + b_{1i})$ over centers is of particular interest in this analysis. Figure 4 shows the density of the prognostic index hazard ratio $HR = \exp(\beta + b_{1i})$ over centers. Considering the 5^{th} and 95^{th} quantiles of this density, the prognostic index hazard ratio lies for 95% of the centers between 1.61 and 2.45 when considering the Allard prognostic index. Figure 5 plots $\exp(b_{0i})$, the predicted center baseline risk, and $\exp(\beta + b_{1i})$, the predicted prognostic index effect.

Discrimination of the Allard prognostic index, when used to separate patients into poor and good prognosis remains good with our data. Our results provide interpretable information for clinicians to decide whether this prognostic index provides hazard ratios which show an acceptable heterogeneity over centers.

5 Simulations

Simulations in Legrand et al. (2005) showed good performance of our approach when using a model similar to (1) to investigate heterogeneity in treatment effect in multicenter clinical trials. The setting for these simulations was similar to the bladder data analysed above but we assumed an equal balance of the two groups defined by the treatment indicator x_{ij}. In the prognostic index setting, however, patients are typically divided into unbalanced prognostic groups. In the following simulations, we will further evaluate the performance of our approach considering various proportions of "good" and "poor" prognostic patients.

To simulate data, values of the baseline event rate $\lambda_0(t)$ (assumed to be constant over time $\lambda_0(t) = \lambda$) and of the treatment effect, β, are chosen to resemble the bladder cancer data set. Considering the same endpoint as in our analysis of the bladder data (DFI), we use a yearly constant baseline event rate of $\lambda = 0.0969$ and a prognostic index effect of $\beta = 0.7372$. This corresponds to a hazard ratio of 2.09 and a 5 year DFI probability of 61.6% and 36.3% in the good and poor prognosis group. The accrual and follow up times are also chosen to resemble the bladder cancer data set. We therefore consider an accrual period of 1065 days (appr. 35 months) and a follow up time of 2440 days (appr. 80 months). Time at risk, rt_{ij}, for a particular patient consists of the time at risk before the end of the accrual period (assuming a constant entry rate over the accrual period) plus the follow up time. This leads to a median follow-up of about 8 years.

We consider the same number of centers, G, and patients per center, $n_i, i = 1, \ldots, G$, as in our real bladder cancer database, namely 35 centers accruing respectively 21, 23, 23, 24, 26, 30, 30, 34, 35, 35, 35, 35, 39, 42, 42, 43, 44, 52, 52, 55, 56, 61, 62, 63, 66, 73, 85, 86, 92, 104, 117, 120, 155, 183, and 249 patients. In each dataset, a proportion p of the patients are randomly assigned to be in the "good" prognosis group, with p taken to be 0.25, 0.50 and 0.75. Different values of the heterogeneity parameters are considered in these simulations, including the case where no heterogeneity is present for one or both random effects.

For each parameter setting, 250 datasets were generated in Splus-2000 from model (1). Given $G = 35$, $\mathbf{n} = (n_1, ..., n_G)$ the size of each center and particular values of the parameters $\lambda, \sigma_0^2, \sigma_1^2$ and β, the observations in a dataset are generated in the following way. First, G random center effects, $b_{01}, \ldots, b_{0G}$ and G interaction random effects $b_{11}, \ldots, b_{1G}$ are independently generated from a zero-mean normal distribution with variance σ_0^2, resp. σ_1^2. The time-to-event outcome for each patient, et_{ij}, is randomly generated from an exponential family distribution with parameter λ_{ij} given by (1) with $\lambda_0(t) = \lambda$. A patient for which the time-to-event is longer than the time at risk is censored with censoring time equal to time at risk so that $t_{ij} = min(rt_{ij}, et_{ij})$ and $\delta_{ij} = I(et_{ij} \leq rt_{ij})$ is the censoring indicator.

For each parameter setting $(G, \mathbf{n}, \lambda, \sigma_0^2, \sigma_1^2, \beta)$, our model is fit using the extended version of The Survival Kit described above, allowing for the joint estimation of the two variance components. In Table 4, we report for each parameter p and for each set of "population parameters" $(\beta, \sigma_0^2, \sigma_1^2)$:

- the bias, computed as the average difference between the estimated values and the true value over the 250 fits,

- the median, computed as the median of the estimated values over the 250 fits,

- the empirical standard deviation computed as the square root of the average squared difference between the estimated value and the mean estimated values over the 250 fits,

- the median model based standard deviation computed as the median over the 250 fits of the estimated standard deviation of β (obtained using standard Cox regression with the estimated random effects as an offset) and of the square root of Var_{θ_0} (resp. Var_{θ_1}) for σ_0^2 (resp. σ_1^2) discussed in Section 3 .

These results indicate that our estimation approach, based on the Laplace approximation, is sufficiently accurate when applied to settings similar to the one we consider whatever the proportion of good and poor prognosis patients. Bias is generally small and the median of the estimated values close to the population parameter. When considering no heterogeneity over centers ($\sigma_0^2 = 0$), our approach appears to underestimate the variance of the random interaction. This is particularily true when considering ($\sigma_1^2 = 0.08$) and a low proportion of patients in the good prognosis group ($p = 0.25$).

The standard deviation (either empirical or model-based) appears to be influenced only to a very small extent by the proportion of patients in each prognosis group. Model-based standard deviations of the variance components should be interpreted, according to the Bayesian paradigm, as parameters characterizing the joint posterior density and are therefore influenced by the skewness of the distribution (Spiegelhalter et al., 2004).

This explains why these values are greater than those obtained for the empirical standard error. Therefore we advise using the whole marginal posterior density rather than the model-based standard deviation for the construction of credible sets.

These simulations demonstrate the good performance of the estimation approach and the ability of our model to adequately identify the source of variation in each of the settings considered. This shows that fitting such a model indeed provides useful information on the heterogeneity in outcome and in risk groups hazard ratios over centers that can be used when studying the validation of a prognostic index.

6 Discussion

Most validation techniques focus on the overall reproducibility of the results in a validation sample. However prognostic indices are practical tools that should be relevant in different centers with possibly different patient populations. It might indeed be the case that the prognostic index risk groups hazard ratio is "on average" the same in the construction and validation set, while important variation exists from center to center, leading to questionable clinical validity of the model. The main idea we defend in this paper is that heterogeneity in the prognostic index risk groups hazard ratios over centers conveys important additional information on the generalisibility of the prognostic index under validation.

To investigate such heterogeneity, we advocate the use of a Cox proportional hazards model including a fixed prognostic index effect, a random center effect and a random interaction between these two factors. We restricted our attention to prognostic indices which divide the patient population into two prognostic groups. Extension to prognostic indices dividing the patient into more than two categories (e.g., "poor", "intermediate" and "good" prognosis) could be performed either by considering the variable representing the prognostic groups as an ordered categorical variable (if we have medical rationale to believe in proportionality of risk over risk groups), or by considering K-1 dummy binary variables to represent the K prognostic groups and introduce them all in the model.

Royston et al. (2004) proposed an "internal-external cross validation" procedure considering data from several independent data sets from studies with the same measured factors.

Their idea was to investigate first whether prognostic discrimination was maintained between the independent studies and second whether the baseline survival distribution was heterogeneous across studies. Rather than using a two-step procedure as proposed by Royston et al. (2004), our model allows to investigate both the variability in the effect of the prognostic index over centers and a possible variability in patient prognosis in the different centers. We believe that the information on heterogeneity in baseline risk and prognostic index risk groups hazard ratio is important as both (as well as their combination) can lead to a conclusion that the prognostic index is not generalisable from one center to the other. Indeed, in case of heterogeneity in prognostic index risk groups hazard ratios over centers, the prognostic index will loose its discriminatory power in part of the centers. And in case of heterogeneity in outcome over centers, the outcome of the patients in the various centers might be so different that it becomes difficult to use the specific prognostic index to define further treatment of the patients (e.g., if the good prognosis group in one center has outcomes similar to the poor prognosis group in another center). It is therefore important to provide tools to investigate wether a prognostic index remains useful despite the heterogeneity in outcome and in prognostic index effect existing between centers.

While Royston and Parmar (2004), considering a fixed effect approach, provides a statistic to test the null hypothesis of no heterogeneity in the prognostic index effect over centers, we propose to quantify this heterogeneity and to interpret it in terms of medically relevant quantities. The strength of our approach is to provide to clinicians clear information with respect to the generalisibility of a particular prognostic index over centers. This information is easily interpretable in the clinical context of a particular prognostic index and can therefore assist them in determining the usefulness of a particular prognostic index.

References

ALLARD, P., BERNARD, P., FRADET, Y. AND TETU, B. (1998) The early clinical course of primary Ta and T1 bladder cancer: a proposed prognostic index. *British Journal of Urology* **81**: 692-698.

ALTMAN, D.G. AND ROYSTON, P. (2000). What do we mean by validating a prognostic model?. *Statistics in Medicine* **19**: 453-473.

BOUFFIOUX, C., DENIS, L., OOSTERLINCK, W., VIGGIANO, G., VERGISON, B., KEUPPENS, F., DE PAUW, M., SYLVESTER, R. AND CHEUVART, B. (1992) Adjuvant chemotherapy of recurrent superficial transitional cell carcinoma: results of a European Organization for Research and Treatment of Cancer randomized trial comparing intravesical instillation of thiotepa, doxorubicin and cisplatin. *Journal of Urology* **148**: 297-301.

BOUFFIOUX, C., KURTH, K.H., BONO, A., OOSTERLINCK, W., KRUGER, C.B., DE PAUW, M. AND SYLVESTER, R. (1995) Intravesical adjuvant chemotherapy for superficial transitional cell bladder carcinoma: Results of two European Organization for Research and Treatment of Cancer randomized trials with mitomycin C and doxorubicin comparing early versus delayed instillations and short-term versus long-term treatment. European Organization for Research and Treatment of Cancer Genitourinary Group. *Journal of Urology* **153** Supplement: 934-941.

COX, D.R. (1972) Regression models in life-tables (with discussion). *Journal of the Royal Statistical Society. Series B* **34**: 187-220.

CRAMER, H. (1971) *Mathematical Methods of Statistics* Princeton: Princeton University Press.

DUCHATEAU, L. AND JANSSEN P. (2005) Understanding heterogeneity in mixed, generalized mixed and frailty models. *The American Statistician* **59**: 143-146.

DUCROCQ, V. AND CASELLA, G. (1996) A Bayesian analysis of mixed survival models. *Genet. Sel. Evol.* 1996; **28**: 509-529.

DUCROCQ, V. AND SÖLKNER, J. (1994) "The Survival Kit", a FORTRAN package for the analysis of survival data. In: 5th World Cong. Genet. Appl. Livest. Prod. **22**: 51-52. Dep. Anim. Poultry. Sci., Univ. of Guelph, Guelph, Ontario, Canada.

DUCROCQ, V. AND SÖLKNER, J. (1998) "The Survival Kit - V3.0", a package for large analysis of survival data. In: 6th World Cong. Genet. Appl. Livest. Prod. **27**: 447-448. Anim. Genetics and Breeding Unit, Univ. of New England, Armidale, Australia.

GLIDDEN, D.V. AND VITTINGHOFF, E. (2004) Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine* **23**: 369-388.

GRAF, E., SCHMOOR, C., SAUERBREI, W. AND SCHUMACHER, M. (1999) Assessment and comparison of prognostic classification schemes for survival data. HARRELL, F.E., LEE, K.L., MARK, D.B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**: 361-387.

IBRAHIM, J.G., CHEN M.H. AND SINHA, D. (2001) *Bayesian survival analysis.* New-York: Springer-Verlag Inc.

JUSTICE, A.C., COVINSKY, K.E. AND BERLIN, J.A. (1999). Assessing the generalisibility of prognostic information. *Annals of Internal Medicine* **130**; 515-524.

KENDALL, M. AND STUART, A. (1977) *The advanced theory of Statistics. Volume 1: Distribution theory. Fourth Edition.* London & High Wycombe, charles Griffin & Company Limited.

KURTH, K.H., SCHRÖDER, F.H., TUNN, U., AY, R., PAVONE-MACALUSO, DEBRUYNE, F., DE PAUW, M., DALESIO, O. AND TEN KATE, F. (1984) Adjuvant treatment of superficial transitional cell bladder carcinoma: preliminary results of an European Organization for Research and Treatment of Cancer randomized trial comparing doxorubicin hydrochloride, ethoglucid and transurethral resection alone. *Journal of Urology* **132**: 258-262.

LEGRAND, C. (2005) Assessing heterogeneity in multicenter clinical trials using the frailty model. *PhD Thesis*; Center for Statistics, Hasselt University.

LEGRAND,C., DUCROCQ, V., JANSSEN, P., SYLVESTER, R. AND DUCHATEAU L. (2005) A Bayesian approach to jointly estimate center and treatment by center heterogeneity in a proportional hazards model. *Statistics in Medicine* **24**: 3789-3804.

MATSUYAMA, Y., SAKAMOTO, J. AND OHASHI, Y. (1998) A Bayesian hierarchical survival model for the institutional effects in a multi-centre cancer clinical trial. *Statistics in Medicine* **17**: 1893-1908.

NELDER, J.A. AND MEAD, R. (1965) A Simplex method for function minimization. *Computer Journal* **7**: 308-13.

NEWLING, D.W., ROBONSON, M.R., SMITH, P.H., BYAR, D., LOCKWOOD, R., STEVENS, I, DE PAUW, M. AND SYLVESTER, R. (1995) Tryptophan metabolites, pyridoxine (vitamin B6) and their influence on the recurrence rate of superficial bladder cancer. Results of a prospective randomized phase III study performed by the EORTC GU Group. EORTC Genitourinary Tract CancerCooperative Group. *European Urology* **27**: 110-116.

OOSTERLINCK, W., KURTH, K.H., SCHRÖDER, F.H., BUTLINCK, J., HAMMOND, B. AND SYLVESTER, R. (1993) A prospective European Organization for Research and Treatment of Cancer Genitourinary Group randomized trial comparing transurethral resection followed by a single intravesical instillation of epirubicin or water in single stage TaT1 papillary carcinoma of the bladder. *Journal of Urology* **149**: 749-752.

RIPATTI, S., LARSEN, K. AND PALMGREN J. (2002) Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. *Lifetime Data Analysis* **8**: 349-360.

ROYSTON, P., PARMAR, M.K.B. AND SYLVESTER, R. (2004) Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine* **23**: 907-926.

SCHEMPER, M. AND STARE, J. (1996) Explained variation in survival analysis. *Statistics in Medicine* **15**: 1999-2012.

SCHEMPER, M. AND HENDERSON, R. (2000) Predicitve accuracy and explained variation in Cox regression. *Biometrics* **56**: 249-255.

SCHEMPER, M. (2003) Predictive accuracy and explained variation. *Statistics in Medicine* **22**: 2299-2308.

SIMON, R. AND ALTMAN, D.G. (1994). Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer* **69**: 979-985.

SINHA, D., IBRAHIN, J.G. AND CHEN, M.H. (2003) A Bayesian justification of Cox's partial likelihood. *Biometrika* **90**: 629-641.

SMITH, A., SKENE, A.M., SHAW, J., NAYLOR, J. AND DRANSFIELD, M. (1985) The implementation of the Bayesian paradigm. *Commun Stat Theor Meth* **14**: 1079-1102.

SMITH, C.T., WILLIAMSON P.R. AND MARSON A.G. (2005) Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine* **24**(9): 1307-1319.

SPIEGELHALTER, D.J., ABRAMS, K., AND MYLES, J.P. (2004) *Bayesian approaches to clinical trials and health care evaluation.* Chichester: Wiley.

SYLVESTER, R.J., VAN DER MEIJDEN, A.P.M., OOSTERLINCK, W., WITJES, J.A., BOUFFIOUX, C., DENIS, L., NEWLING, D.W.W. AND KURTH, K. (2006) Predicting recurrence and progression in individual patients with stage TaT1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORT trials. *European Urology* **49**: 466-477.

TIERNEY, L. AND KADANE, J.B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**(391): 82-86.

VAIDA, F. AND XU, R. (2000) Proportional hazards model with random effects. *Statistics in Medicine* **19**: 3309-3324.

VAN HOUWELINGEN, H.C. (2000). Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* **19**: 3401-3415.

WITJES, J.A., VAN DER MEIJDEN, A.P., SYLVESTER R.C., DEBRUYNE, F.M., VAN AUBEL, A. AND WITJES W.P. (1998) Long term follow up of an EORTC randomized prospecive trial comparing intravesical bacille Calmette-Gurin-RIVM and Mitomycin C in superficial bladder cancer. *Urology* **52**: 403-410.

YAMAGUCHI, T. AND OHASHI, Y. (1999) Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. *Statistics in Medicine* **18**: 1961-1971.

YAMAGUCHI, T., OHASHI, Y. AND MATSUYAMA Y. (2002) Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials. *Statistical Methods in Medical Research* **11**: 221-236.

Table 1: Disease free interval by number of Adverse Tumor Characteristics in the Allard et al. (1998) cohort and in the EORTC cohort

	Allard et al. cohort N=333			EORTC cohort N=2501			
	N (%)	1 year DFI	2 year DFI	N (%)	1 year DFI	2 year DFI	5 year DFI
0 ATC	64 (19.2%)	85.6%	68.5%	411 (16.4%)	83.2%	75.3%	61.6%
1 ATC	97 (29.1%)	66.0%	55.4%	768 (30.7%)	72.9%	60.4%	45.3%
2 ATCs	104 (31.2%)	47.6%	31.7%	761 (30.4%)	61.1%	48.9%	37.6%
3-4 ATCs	68 (20.4%)	29.6%	19.1%	561 (22.4%)	54.8%	41.1%	31.8%

Table 2: Results by PI_{Allard} in the EORTC bladder cancer data.

Prognosis group	Good prognosis group	Poor prognosis group
N (%)	411 (16.4%)	2090 (83.6%)
1 year DFI (95 % CI)	83.2% (79.0-86.6)	63.7% (61.5-65.9)
5 year DFI (95% CI)	61.6% (55.6-67.1)	39.9% (36.3-41.4)
HR (95%CI) / p-value	2.09 (1.73-2.52) / < 0.0001	

Table 3: Estimation of the first three mixed moments obtained either from (a) Gauss-Hermite quadrature using (11)-(13) or from (b) Grid evaluation using (14)-(17)

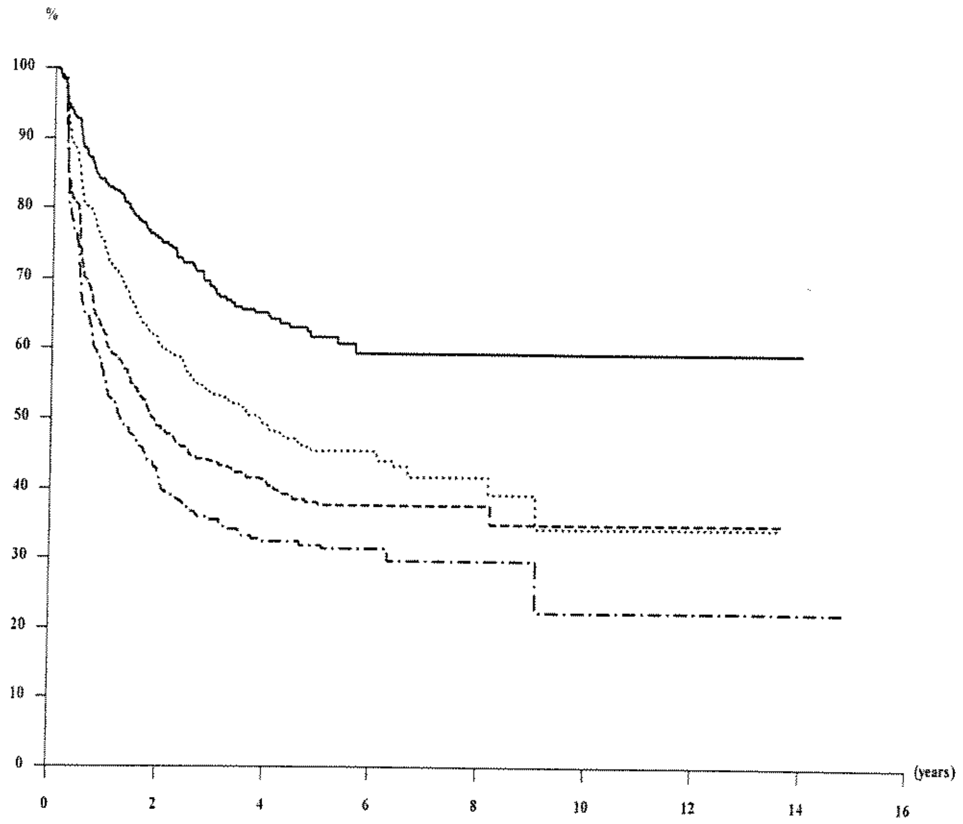
	(a) Gauss-Hermite quadrature			(b) Grid evalutation		
	μ_θ	$\sqrt{V_\theta}$	γ_θ	$\tilde{\mu}_\theta$	$\sqrt{\tilde{V}_\theta}$	$\tilde{\gamma}_\theta$
θ_0	0.0894	0.0497	0.8116	0.0867	0.0415	0.2910
θ_1	0.0534	0.0408	1.1573	0.0520	0.0385	0.9444

Table 4: Simulations. Results for $\beta = 0.7372$, different values of σ_0^2 and σ_1^2 and different values of p. Bias, median value, empirical standard deviation (emp. std) and median model based standard deviation (model std).

	$p = 0.25$			$p = 0.50$			$p = 0.75$		
	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = 0$	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = 0$	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = 0$
Bias	0.0018	0.0021	0.0050	-0.0032	0.0015	0.0022	0.0021	0.0013	0.0014
Median	0.7376	0.0000	0.0000	0.7687	0.0017	0.0023	0.7420	0.0000	0.0000
Emp std	0.0591	0.0038	0.0091	0.0484	0.0030	0.0048	0.0589	0.0032	0.0037
Model std	0.0579	0.0079	0.0246	0.0525	0.0073	0.0127	0.0634	0.0066	0.0080
	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = .04$	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = .04$	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = .04$
Bias	0.0047	0.0024	-0.0035	0.0094	0.0033	-0.0057	0.0077	0.0051	-0.0033
Median	0.7397	0.0000	0.0358	0.7491	0.0000	0.0333	0.7489	0.0000	0.0353
Emp std	0.0695	0.0042	0.0272	0.0695	0.0053	0.0191	0.0767	0.0077	0.0181
Model std	0.0687	0.0092	0.0417	0.0627	0.0108	0.0261	0.0723	0.0149	0.0232
	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = .08$	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = .08$	$\beta = .7372$	$\sigma_0^2 = 0$	$\sigma_1^2 = .08$
Bias	0.0150	0.0026	-0.0203	0.0113	0.0034	-0.0123	0.0046	0.0063	-0.0103
Median	0.7491	0.0000	0.0500	0.7520	0.0000	0.0617	0.7410	0.0000	0.0674
Emp std	0.0769	0.0049	0.0301	0.0677	0.0056	0.0296	0.0855	0.0108	0.0256
Model std	0.0723	0.0097	0.0545	0.0695	0.0123	0.0393	0.0792	0.0182	0.0355

	$p = 0.25$			$p = 0.50$			$p = 0.75$		
	$\beta = .7372$	$\sigma_0^2 = .04$	$\sigma_1^2 = .04$	$\beta = .7372$	$\sigma_0^2 = .04$	$\sigma_1^2 = .04$	$\beta = .7372$	$\sigma_0^2 = .04$	$\sigma_1^2 = .04$
Bias	0.0138	0.0008	-0.0022	0.0028	0.0010	-0.0028	-0.0046	0.0005	0.0026
Median	0.7484	0.0391	0.0310	0.7403	0.0388	0.0342	0.7319	0.0400	0.0408
Emp std	0.0700	0.0182	0.0336	0.0679	0.0203	0.0247	0.0716	0.0246	0.0274
Model std	0.0676	0.0229	0.0453	0.0637	0.0238	0.0336	0.0738	0.0282	0.0329
	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = 0$	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = 0$	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = 0$
Bias	-0.0049	-0.0032	0.0085	0.0016	-0.0030	0.0067	-0.0039	-0.0065	0.0088
Median	0.7286	0.0727	0.0000	0.7418	0.0752	0.0000	0.7367	0.0710	0.0001
Emp std	0.0543	0.0274	0.0168	0.0550	0.0256	0.0110	0.0643	0.0293	0.0137
Model std	0.0587	0.0342	0.0316	0.0537	0.0351	0.0235	0.0648	0.0337	0.0251
	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = .04$	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = .04$	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = .04$
Bias	0.0047	0.0012	0.0017	0.0038	-0.0044	0.0022	-0.0088	0.0028	0.0013
Median	0.7419	0.0812	0.0417	0.7426	0.0726	0.0407	0.7227	0.0768	0.0352
Emp std	0.0687	0.0295	0.0352	0.0648	0.0301	0.0280	0.0725	0.0359	0.0306
Model std	0.0689	0.0363	0.0503	0.0652	0.0355	0.0376	0.0732	0.0405	0.0377
	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = .08$	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = .08$	$\beta = .7372$	$\sigma_0^2 = .08$	$\sigma_1^2 = .08$
Bias	0.0147	0.0005	0.0011	-0.0008	0.0031	-0.0020	-0.0048	-0.0050	0.0029
Median	0.7514	0.0796	0.0762	0.7370	0.0820	0.0724	0.7302	0.0697	0.0797
Emp std	0.0793	0.0277	0.0451	0.0780	0.0334	0.0408	0.0878	0.0377	0.0457
Model std	0.0784	0.0363	0.0614	0.0733	0.0398	0.0496	0.0830	0.0433	0.0507

Figure 1. Bladder cancer data. Disease-free interval probability in 2501 patients with primary or recurrent Ta and T1 bladder cancer (EORTC database) according to the number of adverse tumour characteristics (ATCs) defined by Allard et al. (1998).



O	N	Number of patients at risk :							Allard Risk Group
123	411	215	129	35	6	2	1	0	——— 0 ATC
340	768	339	201	74	18	4	2	0	········ 1 ATC
390	761	260	140	41	15	5	3	0	----- 2 ATC

Figure 2. Joint marginal posterior distribution of θ_0 (center) and θ_1 (interaction) obtained by grid evaluation.

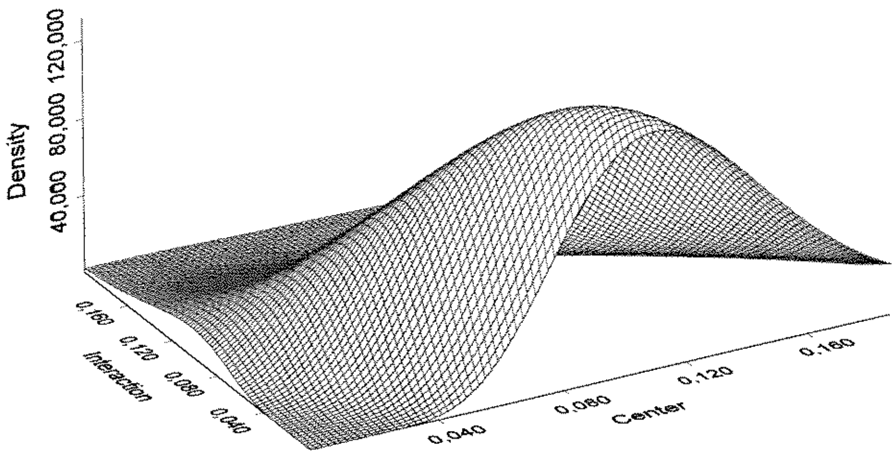


Figure 3. Marginal posterior density of θ_0 (a) and θ_1 (b) obtained from univariate Gram-Charlier approximation based on the first three mixed moments (line) and as obtained from grid evaluation (dots).

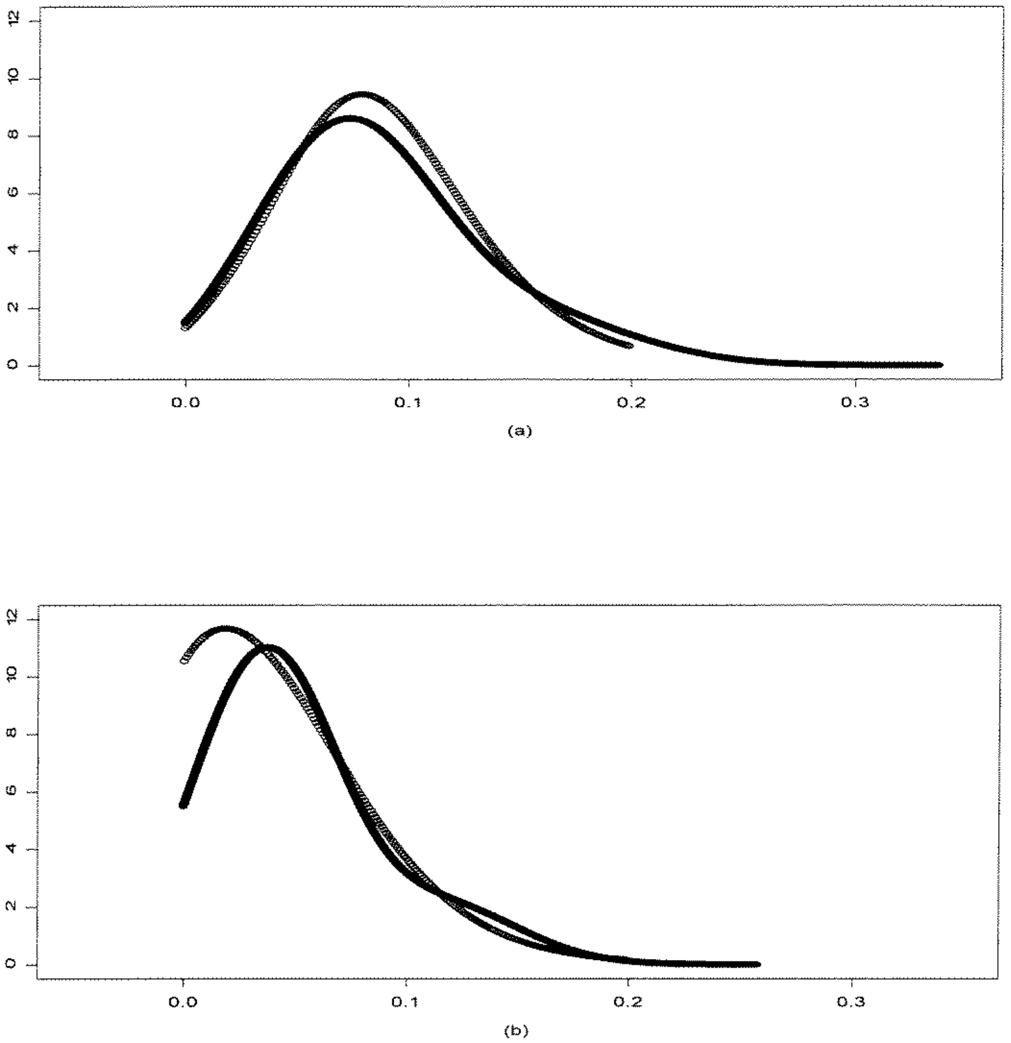


Figure 4. Bladder cancer data. Density of the prognostic index hazard ratio $HR = \exp(\beta + b_{1i})$ over centers.

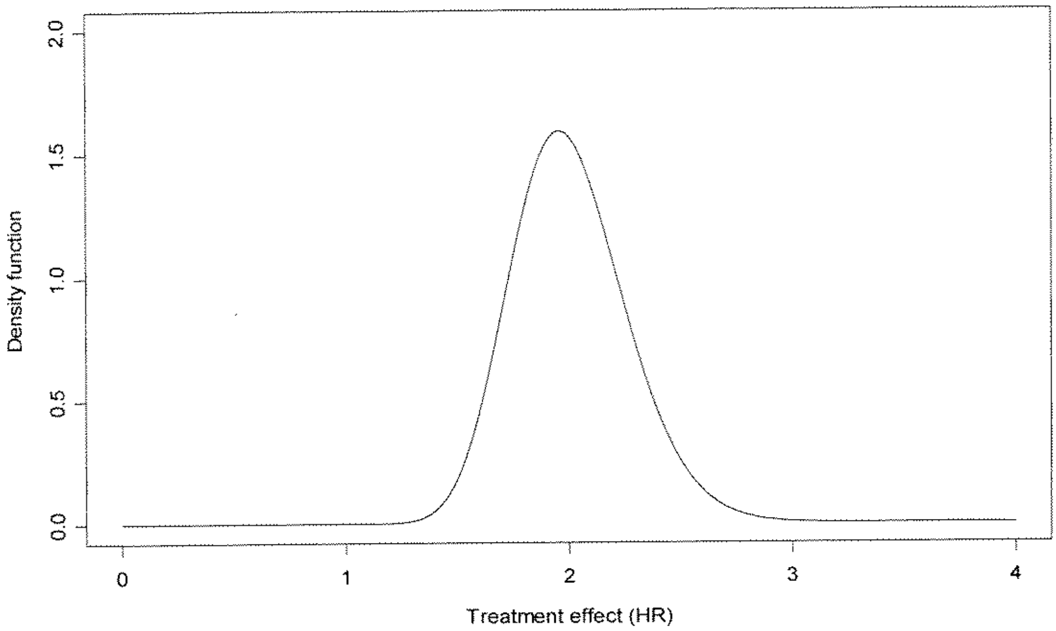
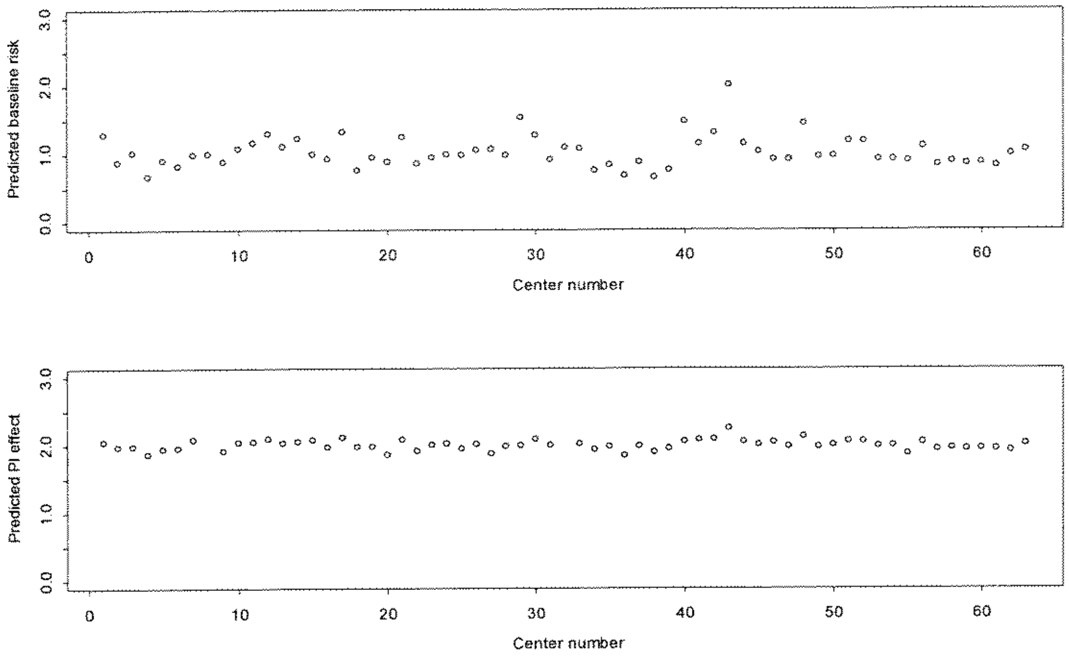


Figure 5. Bladder cancer data. Predicted center baseline risks $(\exp(b_{0i}))$ and predicted prognostic index effects $(\exp(\beta + b_{1i}))$for model (1).



Appendix: Hermite Polynomials

The (probabilists) Hermite polynomials (Cramer, 1971; Kendall and Stuart, 1977) are a polynomial sequence defined by

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}$$

The first several Hermite polynomials are therefore given by

$$\begin{aligned}
H_0(x) &= 1 \\
H_1(x) &= x \\
H_2(x) &= x^2 - 1 \\
H_3(x) &= x^3 - 3x \\
H_4(x) &= x^4 - 6x^2 + 3 \\
&\cdots
\end{aligned}$$

These polynomials are orthogonal with respect to the weight function $w(x) = e^{-x^2/2}$, with

$$\int_{-\infty}^{\infty} H_n(x) H_m(x) e^{-x^2/2} dx = n! \sqrt{2\pi} \delta_{nm}$$

The density function $f_Z(z)$ of any standardised variable Z can be written as (Cramer, 1971)

$$f_Z(z) = \sum_{j=0}^{\infty} (-1)^j \frac{c_j}{j!} H_j(z) \phi(z)$$

To determine the coefficients c_j in this expansion in Hermite polynomials, we multiply this experssion by $H_k(z)$ and integrate from $-\infty$ to ∞ (assuming that the series may be integrated term by term)

$$\int_{-\infty}^{\infty} H_k(z) f_Z(z) dz = \sum_{j=0}^{\infty} \frac{c_j}{j!} \int_{-\infty}^{\infty} H_j(z) H_k(z) \phi(z) dz$$

Using the orthogonality relation, it follows that

$$c_k = (-1)^k \int_{-\infty}^{\infty} H_k(z) f_Z(z) dz$$

From the explicit formulation of the Hermite polynomials, and denoting by μ, σ^2 and γ the mean, variance and skewness of Z, we have (note that $\mu = 0$, $\sigma^2 = 1$)

$$\begin{aligned}
c_0 &= \int_{-\infty}^{\infty} f(z) dz = 1 \\
c_1 &= -\int_{-\infty}^{\infty} z f(z) dz = -\mu = 0 \\
c_2 &= \int_{-\infty}^{\infty} (z^2 - 1) f(z) dz = \sigma^2 - 1 = 0 \\
c_3 &= -\int_{-\infty}^{\infty} (z^3 - 3z) f(z) dz = -\gamma \\
&\ldots
\end{aligned}$$