Estimating the Effect of a Prognostic or Risk Factor after Selection of an "Optimal" Cutpoint

Norbert Holländer^{1,2}, Willi Sauerbrei^{1} and Martin Schumacher^{1,*}

[1] Institute of Medical Biometry and Medical Informatics, University Hospital Freiburg,
Stefan-Meier-Straße 26, D-79104 Freiburg, Germany

[2] Freiburg Center for Data Analysis and Modelling, University of Freiburg,
Eckerstraße 1, D-79104 Freiburg, Germany

* email: ms@imbi.uni-freiburg.de

http://www.imbi.uni-freiburg.de

SUMMARY.

When investigating the effects of potential prognostic or risk factors that have been measured on a quantitative scale, values of these factors are often categorized in two groups. Sometimes an "optimal" cutpoint is chosen that gives the best separation in terms of a two-sample test statistic. If this approach is employed, it leads to a serious inflation of the type I-error. Moreover, the effect of the prognostic or risk factor is overestimated in absolute terms. The problem of inflating the type I-error rate can be avoided when correcting the P-value by using the asymptotic distribution of the maximally selected test statistic or other approximations; the resulting bias of the effect estimates can be reduced by applying shrinkage methods. In this paper, a bootstrap resampling approach is used that yields confidence intervals for the effect of a potential prognostic or risk factor with the desired coverage. The methodology is developed within the framework of the proportional hazards cutpoint model for censored survival data and is illustrated by means of prognostic factor studies in breast cancer.

KEYWORDS:

Maximally test statistics; Cutpoint assessment; P-value correction; Shrinkage methods; Bootstrap resampling; Confidence intervals; Prognostic or risk factors.

Introduction

In the statistical analysis of medical data, almost always some steps of model building or model selection are involved. This may be the choice of a standard method, e.g. a regression model with all factors to be included prespecified or, on the other hand, the use of a complex model-selection strategy within a larger class of candidate models. When analyzing a clinical or epidemiological study we are often in the latter situation although this is not always recognized or, at least, not fully appreciated.

In this paper we will concentrate on a seemingly simple problem of model selection that consists in selecting an "optimal" cutoff-value of a quantitative prognostic or risk factor that gives the best separation between the two resulting groups in terms of a two-sample test statistic. Since this is equivalent to the selection of a cutpoint corresponding to the minimum P-value of the corresponding two-sample test statistic, this approach has also been termed the minimum P-value method (Altman et al. 1994). This problem has attracted some interest both in the statistical and in the medical literature. Starting with the paper by Miller and Siegmund (1982) the statistical properties of various maximally selected test statistics have been worked out (e.g. Koziol 1991; Lausen and Schumacher 1992; Betensky and Rabinowitz 1999; Rabinowitz and Betensky 2000) allowing to calculate the distribution of these test statistics under the null hypothesis that the prognostic or risk factor under consideration does not have any influence on the outcome variable. On the other hand, this approach enjoys some popularity in the medical literature where usually the "optimization" process is not adequately taken into account and simply the minimum P-value is given as the final result of the statistical analysis.

For example, when studying the prognostic relevance of putative prognostic factors in breast cancer, the minimum P-value method has often been used. With regard to investigations on the influence of S-phase fraction and cathepsin-D on disease-free survival time of breast cancer patients several studies were identified where this approach has been taken leading to variety of proposed cutpoints (Altman et al. 1994, Altman 1998). In a recent paper by Linderholm et al. (2000), this approach was used to find a cutpoint for vascular endothelial growth factor (VEGF) content. In the section on statistical methods, the authors write "For the purpose of this study, VEGF was tested as a dichotomous variable and a continuous variable. Survival was estimated using the Kaplan-Meier method, and comparison between study groups was performed with the log-rank test. The optimal cut-off point for VEGF content, according to the lowest P-values and the highest relative risk (RR), was found at 1.75 pg/µg DNA for overall survival, and this was used as the cutpoint in univariate and multivariate analysis". In their analysis, Linderholm et al. (2000) found a significant difference between patients with low VEGF content and those with high VEGF content with an impressive minimum P-value of $P_{min} = 0.0004$. When adjusting for standard prognostic factors in a Cox regression model (Cox 1972) the relative risk of death for VEGF using 1.75 pg/µg DNA as cutpoint was estimated as 1.82 with a 95%-confidence interval [1.11 ; 2.97]; the P-value still being $P_{min} = 0.0170$. The authors concluded that VEGF content is a predictor of overall survival in primary node-positive breast cancer.

The questions that arise from such an extensive use of cutpoint selection are self-evident: Is the reported P-value correct? Is the estimated relative risk and given confidence interval reliable? Are the conclusions valid? Whereas the correction of P-values and the reduction of resulting bias of effect estimates has been dealt with in previous papers (Lausen and Schumacher, 1992; Altman et al. 1994; Schulgen et al. 1994; Schumacher, Holländer and Sauerbrei, 1997) the validity of confidence intervals has not been considered to that extent so far. We show that the naive calculation of confidence intervals suffers from similar defects as P-values and effect estimates; thus a bootstrap resampling approach is proposed. Together with a shrinked estimate the bootstrap based confidence interval yields the desired coverage.

Study on prognostic value of S-phase fraction

The database of the study consists of all patients with primary, previously untreated node positive breast cancer who were operated between 1982 and 1987 in the Department of Gynecology at the University of Freiburg and whose tumor material was available for DNA investigations. Some exclusion criteria (history of malignoma, T_4 and / or M_1 tumors according to the TNM classification system, without adjuvant therapy after primary surgery, older than 80 years etc.) were defined retrospectively but before the statistical analysis. This left 139 patients out of 218 originally investigated for the analysis.

Eight patients characteristics were investigated. Besides the documentation of standard prognostic factors in node positive breast cancer DNA flow cytometry was used to measure ploidy status of the tumor and S-phase fraction, which is the percentage of tumor cells in the DNA synthetizing phase obtained by cell cycle analysis. In the sequel, we consider only S-phase fraction as a potential prognostic factor that was available in 109 patients. The median follow-up was 83 months. The endpoint considered is event-free survival which is defined as the time from surgery to the first of the following events: occurrence of locoregional recurrence, distant metastasis, second malignancy or death. Event-free survival was estimated as 50% after five years. Further details of the study, in the sequel referred to as the Freiburg DNA study, can be found elsewhere (Pfisterer et al. 1995); the data have been used previously (Altman et al., 1994; Schumacher et al., 1997) and are published in Lausen and Schumacher (1996).

P-values and confidence intervals in the proportional hazards cutpoint model

In the sequel, we restrict ourselves to the problem of selecting only one cutpoint and to a so-called univariate analysis. This means that we consider only one covariate X ? in the Freiburg DNA breast cancer data the S-phase fraction ? as a potential prognostic factor. If this covariate has been measured on a quantitative scale the proportional hazards (Cox, 1972) cutpoint model is defined as

$$\lambda\left(t \mid X > \mu\right) = \exp(\beta)\,\lambda\,(t|X \le \mu), \quad t > 0 \qquad (1)$$

where $\lambda\left(t \mid \cdot\right) = \lim_{h \to 0}\,(1/h\,\Pr\left(t \le T < t + h \mid T \ge t, \cdot\right)$ denotes the hazard function of the event-free survival time random variable T. The parameter $\theta = \exp\left(\beta\right)$ is referred to as the relative risk of observations with $X > \mu$ with respect to observations with $X \le \mu$ and is estimated through $\hat{\theta} = \exp\left(\hat{\beta}\right)$ by maximizing the corresponding partial likelihood (Cox, 1972) with given cutpoint μ. The minimum P-value method is a data-dependent categorization method where ? within a certain range of the distribution of X, the selection interval ? the cutpoint $\hat{\mu}$ is chosen so that the P-value for the comparison of observations below and above the cutpoint is a minimum. In the proportional hazards cutpoint model, the logrank test (Peto and Peto 1972) or the Wald or likelihood ratio test derived from the partial likelihood is used for those comparisons. A confidence interval for β is then calculated as

$$\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)} \qquad (2)$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution and $\widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)$ is the model-based estimated variance of $\hat{\beta}$ derived from the proportional hazards cutpoint model (1) as if the estimated cutpoint $\hat{\mu}$ was known in advance.

Due the optimization process involved it is obvious that the minimum P-value method cannot lead to correct results of the logrank test. However, this problem can be solved by using a corrected P-value p_{cor} as proposed in Lausen and Schumacher (1992), which has been developed by generalizing an earlier result of Miller and Siegmund (1982). The formula reads

$$p_{\text{cor}} = \varphi(z)\left[z - \frac{1}{z}\right]\log\left[\frac{(1-\varepsilon)^2}{\varepsilon^2}\right] + 4\frac{\varphi(z)}{z} \qquad (3)$$

where f denotes the standard normal probability density function and z is the (1 – p_{min} / 2)-quantile of the standard normal distribution. The selection interval is characterized by the proportion e of smallest and largest values of X that are not considered as potential cutpoints. It should be mentioned that other approaches of correcting the minimum P-value could be applied; a comparison of three approaches can be found in a paper by Hilsenbeck and Clark (1996). Especially, if there are only a few number of cutpoints an improved Bonferroni inequality can be applied (Worsley, 1982; Lausen, Sauerbrei and Schumacher, 1994; Lausen and Schumacher, 1996).

In order to correct for overestimation it has been proposed (Verweij and Van Houwelingen, 1993) to shrink the parameter estimates by a so called shrinkage factor. Considering the cutpoint model the log-relative risk is then estimated by

$$\hat{\beta}_{\text{cor}} = c \cdot \hat{\beta} \qquad (4)$$

where $\hat{\beta}$ is based on the minimum P-value method and c is the estimated shrinkage factor. Values of c close to one should indicate a minor degree of overestimation whereas small values of c should reflect a substantial overestimation of the log-relative risk. For the estimation of c it has to be considered that maximum partial likelihood estimation in a model

$$\lambda\left(t \mid X > \mu\right) = \exp\left(c\hat{\beta}\right)\lambda\left(t \mid X \le \mu\right) \qquad (5)$$

yields $\hat{c} = 1$ since $\hat{\beta}$ is already the maximum partial likelihood estimate. Thus cross-validation and resampling approaches have to be employed to get sensible estimates of the shrinkage factor; see Schumacher et al. (1997) for a comparison of several methods. In this paper a so-called heuristic estimate

$$\hat{c} = \left(\hat{\beta}^{2} - \widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)\right) \Big/ \hat{\beta}^{2} \qquad (6)$$

was applied where $\hat{\beta}$ and $\widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)$ are defined as above (Van Houwelingen and Le Cessie, 1990). If $\hat{c}$ happened to take negative values, it was set equal to zero. This heuristic estimate performed quite well when compared to more elaborated cross-validation and resampling approaches of estimating the shrinkage factor c (Schumacher et al., 1997) and was taken here for reasons of simplicity. A confidence interval for β is then calculated as

$$\hat{c}\hat{\beta} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)}. \qquad (7)$$

The confidence intervals calculated so far cannot be expected to show the desired properties with regard to coverage. The reason for this is that the variance of the estimated log-relative risk is still model-based, i.e. is derived from a proportional hazards cutpoint model where a fixed and predefined cutpoint is assumed.

In order to take the additional variability of both the estimated cutpoint and the estimated shrinkage factor $\hat{c}$ into account, here a new approach based on bootstrap resampling is used. For that, the complete patient's vector (survival time, censoring indicator, covariate) is sampled with replacement, the sample size being the same as in the original data. This process is repeated B times resulting in B bootstrap samples. In each of these a cutpoint is estimated by the minimum P-value method, the log-relative risk and a shrinkage factor (6) is estimated. The empirical variance of the shrinked log-relative risk over the bootstrap samples is then obtained as

$$\widehat{\text{var}}_{\text{boot}}\left(\hat{c}\hat{\beta}\right) = \frac{1}{B-1} \sum_{j=1}^{B} \left(\hat{c}_j \hat{\beta}_j - \overline{c\beta}_{\text{boot}}\right)^2 \qquad (8)$$

where $\overline{c\beta}_{\text{boot}}$ denotes the average of the shrinked estimated log-relative risks over the B bootstrap samples.

For the calculation of a confidence interval for β the model-based variance in formula (7) is replaced by the empirical bootstrap variance (8).

When applying the minimum P-value method to the data on S-phase fraction, we obtain, based on the logrank test, a selected cutpoint of 10.7% and a minimum P-value of p_{min} = 0.007 when the range between the 10%- and the 90%-quantile of the distribution of S-phase values is used as selection interval. The difference in event-free survival between the two resulting groups is rather impressive and is reflected by an estimated log-relative risk of 0.864. The corresponding 95%-confidence interval of [0.239 ; 1.489] indicates an effect of S-phase fraction on event-free survival. After the P-value correction (3) we obtain p_{cor} = 0.123 clearly providing no indication that S-phase is of prognostic relevance. The correction of the log-relative risk estimate by applying the shrinkage factor (6) leads to a somewhat smaller value of 0.743. Using the naive, model-based variance leads to a corresponding 95%-confidence interval of [0.118 ; 1.368] that does not include a log-relative risk equal to 0; it is therefore not consistent with the corrected P-value of p_{cor} = 0.123.

When using the bootstrap approach with B = 100 bootstrap samples we obtain a confidence interval that is much wider as compared to the others and does not contain the value $\beta = 0$ and is therefore consistent with the corrected P-value of $p_{\text{cor}} = 0.123$. Table 1 summarizes the results of the study in terms of log-relative risks and relative risks.

Table 1: Risk estimates with corresponding 95% confidence intervals [,] for S-phase fraction based on optimal cutpoint in the Freiburg DNA study

	log-relative risk		relative risk	
$\hat{\beta},\ \widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)$	0.864	[0.239 ; 1.489]	2.37	[1.270 ; 4.433]
$\hat{c}\hat{\beta},\ \widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)$	0.743	[0.118 ; 1.368]	2.10	[1.125 ; 3.927]
$\hat{c}\hat{\beta},\ \widehat{\text{var}}_{\text{boot}}\left(\hat{c}\hat{\beta}\right)$ *	0.743	[-0.300 ; 1.786]	2.10	[0.741 ; 5.966]

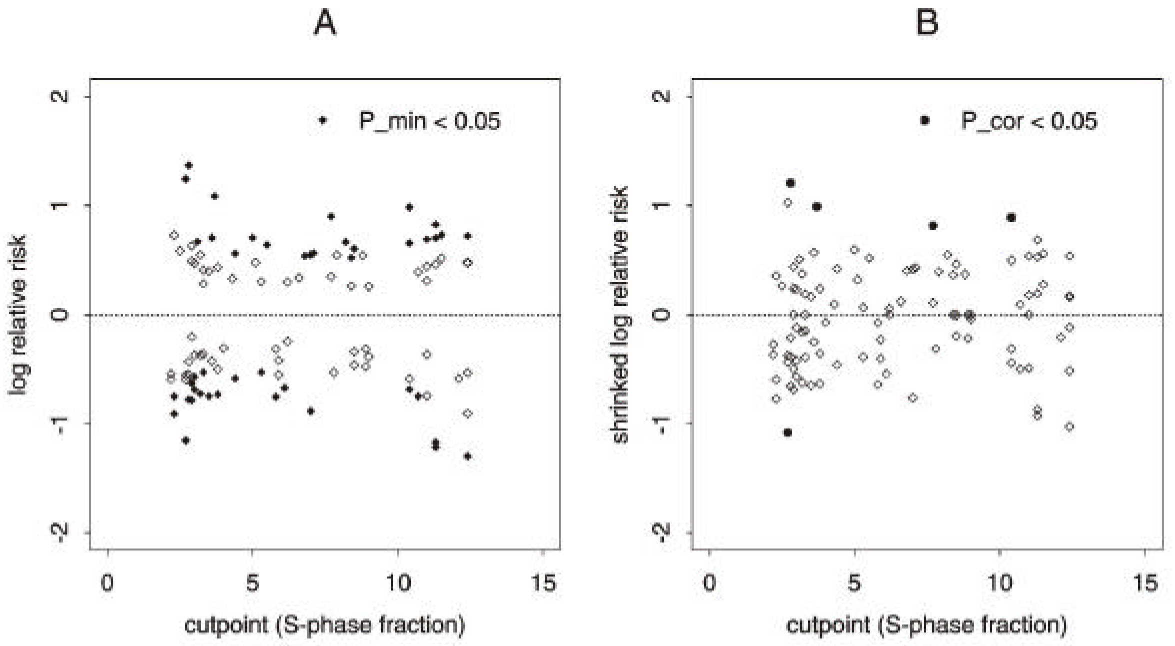
* $\widehat{\text{var}}_{\text{boot}}\left(\hat{c}\hat{\beta}\right)$ based on B = 100 bootstrap samples of the original data

Random re-allocation of S-phase values

By simulating the null hypothesis of no prognostic relevance of SPF with respect to event-free survival we illustrate that the minimum P-value method may lead to a drastic overestimation of the absolute value of the log-relative risk. By a random re-allocation of the observed values of SPF to the observed survival times we simulate independence of these two variables, which is equivalent to the null hypothesis that the log-relative risk associated with S-phase fraction, denoted by β, is equal to zero. This procedure was repeated a hundred times and in each repetition we selected a cutpoint by using the minimum P-value method. In the hundred repetitions, we obtained 44 significant $\left(p_{min} < 0.05\right)$ results for the logrank test corresponding well to theoretical results as outlined in Lausen and Schumacher (1992).

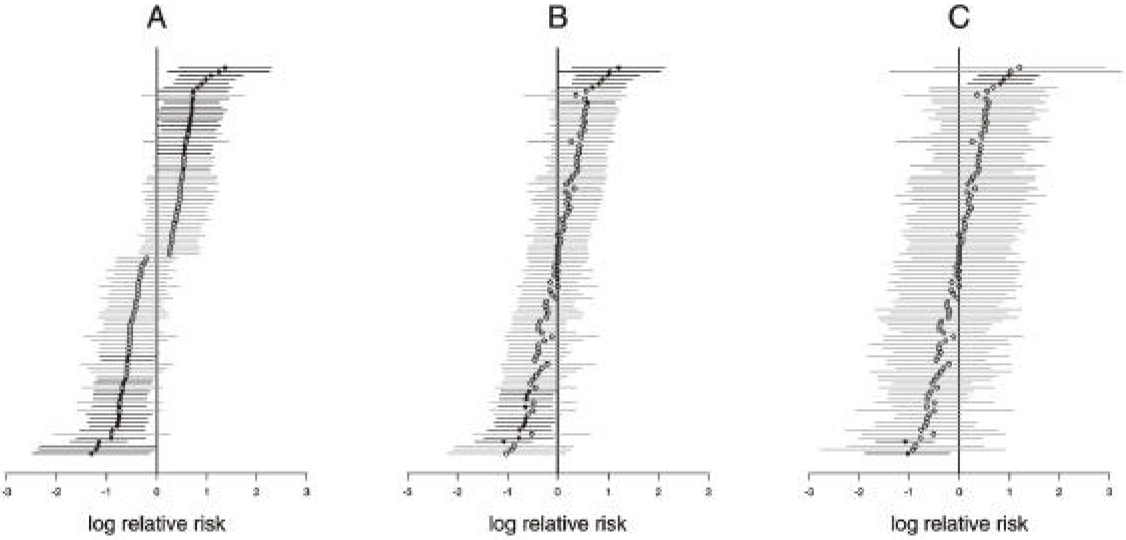
The estimated 'optimal' cutpoints of the hundred repetitions and the corresponding estimates of the log-relative risk are shown in figure 1A. We obtained no estimates near the null hypothesis $\beta = 0$ as a result of the optimization process of the minimum P-value approach. Using the correction formula in the hundred repetitions of the sampling experiment we obtained 5 significant results $\left(p_{cor} < 0.05\right)$ corresponding well to the significance level of $\alpha = 0.05$. Figure 1B shows the results of the correction process where in addition to the correction of P-values, the log-relative risks have been corrected by applying the shrinkage factor (6). It can be seen that we now obtain also values close to the null hypothesis $\beta = 0$; thus the bias induced by model selection has been reduced considerably.

Figure 1: Estimated "optimal" cutpoints and log-relative risks in 100 repetitions of randomly re-allocated S-phase values to event-free survival times before (A) and after (B) correction.



In order to demonstrate that not only the P-value but also confidence intervals are not valid we use again the sampling experiment described above. The results of 100 repetitions in terms of confidence intervals are displayed in figure 2A. It is seen that 39 of these confidence intervals do not contain the value of $\beta = 0$ corresponding well to the number of significant results according to the minimum P-value. Figure 2B shows the resulting confidence intervals after shrinkage of the estimated log-relative risk has been applied. There are still 17 intervals that do not contain the value of $\beta = 0$; so the desired coverage is not obtained through the shrinkage of estimates. However, when using the empirical variance of the shrinked estimates $\hat{c}\hat{\beta}$ in 100 bootstrap samples in each repetition for the calculation of the confidence intervals (Figure 2C) we end up with only 5 intervals that do not contain the value of $\beta = 0$. As can be seen the width of these confidence intervals is considerably larger than that of those where the naive, model based variance of the estimated log-relative risk is used.

Figure 2: 95%-confidence intervals for log-relative risk of S-phase fraction in 100 repetitions of randomly re-allocated S-phase values to event-free survival times (A: estimate based on "optimal" cutpoints, naive model-based variance; B: shrinked estimate, naive model-based variance; C: shrinked estimate, empirical variance from 100 bootstrap samples); confidence intervals not including $\beta = 0$ are marked with filled circles, the samples are ordered according to the values of $\hat{\beta}$ from smallest to largest.



Simulation study

The random re-allocation experiment is based on one specific data set and gives only results that are valid under the null-hypothesis '$\beta = 0$" . In order to investigate whether these results hold more generally a simulation study was performed; we considered one covariate X taken as uniformly distributed on the interval (0 , 1), and, for simplicity, we restricted ourselves to the situation of no censoring. It is assumed that there is a true cutpoint $\mu = 0.5$. Assuming β ranging from 0 to 1 in steps of 0.1 - corresponding to relative risks ranging from 1 to 2.72 - the survival time random variable T has been taken from an exponential distribution with parameter $\lambda = 1$ for $X \le \mu$ and $\lambda = \exp(\beta)$ for $X > \mu$ according to the proportional hazards cutpoint model (1). The choice of the values of β implies that large values of X are associated with a higher risk than for small values of X. This, however, is not a restriction since one can simply switch from X to X - 1. In the absence of censoring we can also assume $\lambda = 1$ without loss of generality. For each parameter constellation, we used

1000 simulated data sets with n = 100 patients.

In a previous publication, Schumacher et al. (1997) have shown that, within the range of values considered, the log-relative risks are overestimated in absolute terms. This problem is particular prominent for small and moderate effects, i.e. values of β ranging from 0 to 0.5 corresponding to relative risks between 1 and 1.6, whereas for very large values of β some underestimation has been observed. In addition, we were able to show that the application of shrinkage can reduce the resulting bias considerably and, in particular, that the heuristic shrinkage factor (6) compared quite well to other, more elaborated methods based on cross-validation and resampling. Table 2 shows the estimated coverage of confidence intervals for the log-relative risk where the nominal level has been set to 95%. For $\beta = 0$ to $\beta = 0.3$ – the latter corresponding to a small effect – it can be seen that the coverage is far away from the desired one and shrinkage alone is not able to compensate for this. However, taking the empirical variance from only B = 20 bootstrap samples leads to a coverage of more than 90%. For large effects ($\beta = 0.6$ to $\beta = 0.8$) model building does not affect the coverage at all whereas $\beta = 0.4$ to $\beta = 0.5$ representing moderate effects is somewhat in the middle of the two extreme situations.

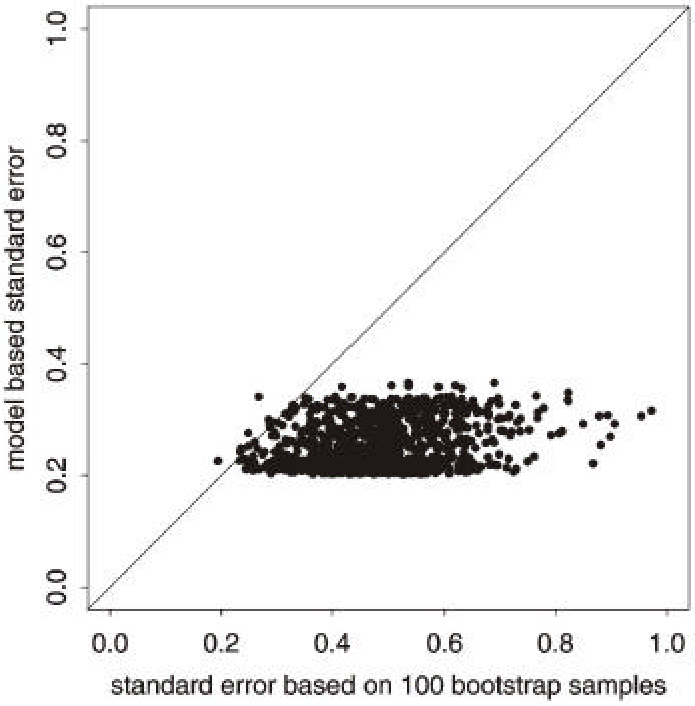
Table 2: Coverage (in percent) of confidence intervals for log-relative risk estimated from 1000 simulated data sets (n = 100); nominal confidence level 95%

	β=0	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5	β=0.6	β=0.7	β=0.8
$\hat{\beta},\ \widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)$	59.5	59.7	64.3	75.2	84.1	94.4	97.5	98.0	98.5
$\hat{c}\hat{\beta},\ \widehat{\text{var}}_{\text{mod}}\left(\hat{\beta}\right)$	76.9	79.3	80.9	86.5	89.5	96.3	97.4	97.1	97.4
$\hat{c}\hat{\beta},\ \widehat{\text{var}}_{\text{boot}}\left(\hat{c}\hat{\beta}\right)$*	91.7	92.1	90.6	91.3	90.2	93.1	95.1	97.3	98.8

* $\widehat{\text{var}}_{\text{boot}}\left(\hat{c}\hat{\beta}\right)$ based on B = 20 bootstrap samples of each simulated data set

In order to investigate whether the coverage could be further improved by increasing the number of bootstrap samples we performed a second simulation study following the same design as the first but only using values $\beta = 0$ and $\beta = 0.5$. For the estimate based on the "optimal" cutpoint and the shrinked estimate, in combination with the model-based variance when calculating the confidence intervals, we obtained results comparable to those already displayed in table 2. For the bootstrap confidence interval based on the shrinked estimate and on the empirical variance from B = 100 bootstrap samples we obtained a coverage of 95.2% for $\beta = 0$ and of 96% for $\beta = 0.5$, respectively. Thus the desired coverage was achieved when a sufficient number of bootstrap replications was used. Figure 3 shows a comparison of standard errors derived from the model-based variance with those derived from the empirical variance based on B = 100 bootstrap samples for the situation $\beta = 0$. In this situation, the model-based variance should approximately be equal to $\left(1/ n_\mu + 1/ \left(n - n_\mu \right) \right)$ where n_μ is the number of patients with $X \leq \mu$ (Kalbfleisch and Prentice 1980). Since the selection interval is taken as the range between the 10%- and 90%-quantile of the distribution of X, the standard errors derived from the model-based variance should range between 0.2 ($n_\mu = 50$) and 0.33 ($n_\mu = 10$ or $n_\mu = 90$). It is clearly seen that these standard errors are much smaller than those based on the bootstrap variance in most of the simulated data sets.

Figure 3: Standard errors of log-relative risk derived from naive model-based variance vs. those based on the empirical variance of 100 bootstrap samples; results are given for 1000 simulated data sets with $\beta = 0$.



In this second simulation study, we also compared the type I error rate and the power when the test decision would be based on the various P-values and confidence intervals. Table 3 summarizes the results. It can be seen that test decisions based on confidence intervals using the model based variance agree well with the use of minimum P-values whereas the bootstrap confidence intervals reflect very closely the test decisions that would have been based on the corrected P-values.

Table 3: Test decisions (in percentage) based on confidence intervals for log-relative risk and on P-values estimated from 1000 simulated data sets (n = 100); nominal significance level 5%

	$\beta = 0$		$\beta = 0.5$	
	$0 \notin$ CI	$P < 0.05$	$0 \notin$ CI	$P < 0.05$

estimate based on "optimal" cutpoint;
naive model-based variance;
minimum P-value
	42.0	43.2	90.4	90.8

shrinked estimate; naive model-based variance	19.8		74.0	
shrinked estimate; empirical variance from 100 bootstrap samples; corrected P-value	4.8	6.7	40.4	44.9

Discussion

In studies investigating the effects of potential prognostic or risk factors, values of the factors considered are often categorized into two or three categories. Sometimes this may be done according to medical or biological reasons or may just reflect some consensus in the scientific community. When a "new" prognostic or risk factor is investigated the choice of such a categorization represented by one or more cutpoints is by no means obvious. Thus, often an attempt is made to derive such cutpoints from the data and to take those cutpoints that give the best separation in the data at hand.

In general the minimum P-value method leads to a dramatic inflation of the type I-error rate; the chance of declaring a quantitative factor as prognostically relevant when in fact it does not have any influence on event-free survival, is about 50% when a level of 5% has been intended. Thus, correction of P-values is essential but leaves the problem of overestimation of the relative risk in absolute terms. The latter problem is especially relevant when sample sizes and / or effect sizes are of small or moderate magnitude and can, at least partially, be solved by applying some shrinkage method. The construction of valid confidence intervals, however, is still on open problem. We have proposed a bootstrap approach that has two essential ingredients. The first is the application of a shrinkage factor to the estimated log-relative risk in order to reduce the bias induced by selecting an "optimal" cutpoint. The second ingredient consists of repeating the whole model selection process employed within each bootstrap sample, i.e. selection of an "optimal" cutpoint, estimation of the log-relative risk based on that cutpoint and application of a shrinkage factor. The resulting empirical variance of the shrinked estimated log-relative risks over the bootstrap samples is then capable to capture all sources of variability and thus provides a valid basis for the construction of confidence intervals. In contrast, the use of the naive model-based variance is strictly not valid if the cutpoint has not been prespecified in advance.

In two simulation studies we have shown that the desired coverage of the confidence intervals can be obtained even with a moderate number of bootstrap samples, say 100. In addition, the test decisions based on the bootstrap confidence intervals are in agreement with the corrected P-values although the two methods do not yield identical results. So in the random re-allocation experiment, for example, only four of the five replications with $P_{cor} < 0.05$ are identical to those where the value of $\beta = 0$ is not contained in the bootstrap confidence intervals. One possible reason is that the confidence intervals are closely related to Wald-type tests whereas the correction of P-values has been applied to the logrank test as a generalized rank test (Peto and Peto 1972) that can be derived as the score test from the partial likelihood in the proportional hazards cutpoint model (1).

It should be noted, however, that the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. (1994) point out this problem for studies of the prognostic relevance of S-phase fraction in breast cancer published in the literature; they identified 19 different cutpoints used in the literature, some of them were solely used because they emerged as the "optimal" cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients (Ferrandina et al. 1997) 12 studi es were included with 12 different cutpoints thus adding an additional source of heterogeneity to this meta-analysis (Altman 2001). Interestingly, neither cathepsin-D nor S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of the American Society of Clinical Oncology (Bast et al. 2001).

Often an "optimal" cutpoint is determined by the minimum P-value method by a univariate analysis; i.e. by taking only this prognostic or risk factor into account, which is then also used in a multivariate analysis by including the resulting binary variable together with other variables in a Cox regression model. By doing so, the bias from the univariate analysis is transferred to the multivariate setting (Altman et al. 1994; Altman 1998). In the paper by Linderholm et al. (2000) cited in the introduction, vascular endothelial growth factor (VEGF) content is investigated together with tumor size, histopathologic grade, number of lymph nodes and estrogen receptor status in a Cox regression model. VEGF had still a significant effect (P_{min} = 0.017) on overall survival; the log-relative risk is estimated as 0.599 with a 95%-confidence interval [0.104 ; 1.089]. Since we do not have access to the original data of their study, we cannot apply the methodology outlined in this paper. However, by applying the simple formula given by Altman et al. (1994) we can conclude that the corrected P-value would be larger than 0.05 when it is assumed that for the selection of cutpoints all values between the 10%- and 90%-quantile have been considered. The standard error based on the naive model-based variance is 0.251 and thus the resulting estimated shrinkage factor (6) for the log-relative risk is 0.824. Thus it remains at least questionable whether VEGF content is of prognostic relevance in breast cancer.

Some colleagues suggested that it would be desirable to have a simple formula at hand to correct published confidence intervals in a similar way as it is possible with the P-value correction (Altman et al., 1994) mentioned above. When the model-based variance is available the shrinkage factor (6) and a shrinked log-relative risk can be calculated. A confidence interval can be derived in the situation that the null-hypothesis "$\beta = 0$" holds true by taking the $\left(1-\alpha/2\right)$-quantile of the asymptotic distribution of the maximally selected logrank statistic instead of the corresponding one of the standard normal distribution. For α = 0.05 this would be 3.054 instead of 1.96 when taking the range between the 10%- and 90%-quantile as the selection interval (Lausen and Schumacher, 1992). In the Linderholm et al. (2000) paper this procedure would yield a 95%-confidence interval of [-0.272 ; 1.261] for the log-relative risk. It should be stressed, however, that this procedure is not generally applicable since it would lead to confidence intervals that are for too wide when the true effects are of moderate or large size. This corresponds to experience that we have already made before when investigating various strategies for the shrinkage of effect estimates (Schumacher et al., 1997) in order to reduce the bias induced by the model selection process.

With regard to shrinkage, we used the so-called heuristic shrinkage factor (6) throughout this paper. In the definition of this factor the naive model-based variance is used that we showed to be invalid for the calculation of confidence intervals. In order to resolve this contradiction we compared the heuristic shrinkage factor with a shrinkage factor based on an additive bootstrap approach; see Schumacher et al. (1997) for details. In this further investigation (data not shown in detail) it turned out that the differences were negligible thus favouring the heuristic shrinkage factor because of its simplicity.

A key characteristic of the approach taken in this paper is that, at the end, there is one so-called "final model". In our setting, this is an estimated cutpoint, an estimated log-relative risk and a valid confidence interval for the log-relative risk. There are recent, alternative approaches taking the model selection uncertainty explicitely into account. One of these approaches is Breiman's bagging (Breiman 1996) that - applied to the cutpoint problem - results in an estimated log-relative risk as a function of the potential cutpoints derived from bootstrap resampling (Schumacher, Holländer and Sauerbrei 1996). In the bagging approach, repetition of the whole model selection process within each bootstrap sample including shrinkage is essential. Another approach would be Bayesian model averaging where the posterior distribution of both cutpoint and log-relative risk needs to be determined (Chatfield 1995, Draper 1995, Volinsky et al. 1997, Hoeting et al. 1999).

We have studied the cutpoint model mainly for two reasons: the first is its prominence in the medical as well as in the statistical literature; to the considerations already outlined above we only mention that the data-driven choice of cutpoints is the basic building-stone of the well-known classification and regression trees (CART) approach (Breiman et al. 1984). The second reason is that we consider the cutpoint problem as some kind of prototype for other, more complicated problems of model selection as the well-known variable selection problem (e.g. Miller 1990), selection of the functional form of effects of quantitative factors (e.g. Royston and Altman 1994), searching for interactive effects etc. Admittedly, these problems are much more complex to allow a straight forward generalization of results presented in this paper but the general idea of repeating the whole model selection process including bias correction within each bootstrap sample to obtain valid confidence intervals based on bootstrap resampling seems promising also in other settings (Hjorth 1994, Sauerbrei 1999).

References

Altman, D. G. (1998). Suboptimal analysis using "optimal" cutpoints. *British Journal of Cancer* **78**, 556-557.

Altman, D. G. (2001). Systematic reviews of evaluation of prognostic variables. In: Egger M, Smith GD, Altman DG, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London

Altman, D. G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994). Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* **86**, 829-835.

Bast, R.C., Ravdin, P., Hayes, D.F., Bates, S. Fritsche, H., Jessup, J.M., Kemeny, N., Locker, G.Y., Mennel, R.G. and Somerfield, M.R. for the American Society of Clinical Oncology Tumor Markers Expert Panel (2001). 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: Clinical practice guidelines of the American Society of Clinical Oncology. *Journal of Clinical Oncology* **19**, 1865-1878.

Betensky, P. and Rabinowitz, D. (1999). Maximally selected chi-square statistics for $k \times 2$ tables. *Biometrics* **55**, 317-320.

Breiman, L., Friedman, J. H., Olsen, R. A. and Stone, C. J. (1984*). Classification and Regression Trees*. Wadsworth, Monterey.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **26**, 123-140.

Chatfield, C. (1995). Model selection uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society (Series A)* **158**, 419-466.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, Series B **74**, 187-200.

Draper, D. (1995). Assessment and propagation of model selection uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)* **57**, 45-97.

Ferrandina, G., Scambia, G. Bardelli, F., Benedetti Panici, P., Mancuso, S. and Messori, A. (1997). Relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients: a meta-analysis. *British Journal of Cancer* **76**, 661-666.

Halpern, A. L. (1999). Minimally selected p and other tests for a single abrupt changepoint in a binary sequence. *Biometrics* **55**, 1044-1050.

Hilsenbeck S. G. and Clark G. M. (1996). Practical P-value adjustment for optimally selected cutpoints. *Statistics in Medicine* **15,** 103-112.

Hjorth, J.S.U. (1994). Computer intensive statistical methods. Validation, model selection and bootstrap. London: Chapman and Hall.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science* **14**, 382-417.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Koziol, J. A. (1991). On maximally selected chi-square statistics. *Biometrics* **47**, 1557-1561.

Lausen B., Sauerbrei W. and Schumacher M. (1994). Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In: Dirschedl P, Ostermann R, eds. *Computational Statistics*. Heidelberg: Physica-Verlag, 1483-1496

Lausen, B. and Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics* **48**, 73-85.

Lausen, B. and Schumacher, M. (1996). Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics & Data Analysis* **21**, 307-326.

Linderholm, B., Graukvist, K., Wilking, N., Johansson, M., Tavelin, B. and Henriksson, R. (2000). Correlation of vascular endothelial growth factor content with recurrences, survival, and first relapse site in primary node-positive breast carcinoma after adjuvant treatment. *Journal of Clinical Oncology* **18**, 1423-1431.

Miller A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

Miller, R. and Siegmund, D. (1982). Maximally selected chi-square statistics. *Biometrics* **38**, 1011-1016.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society (Series A)* **135**, 185-198.

Pfisterer, J., Kommoss, F., Sauerbrei, W., Menzel, D., Kiechle, M., Giese, E., Hilgarth, M. and Pfleiderer, A. (1995). DNA flow cytometry in node positive breast cancer: prognostic value and correlation to morphological and clinical factors. *Analytical and Quantitative Cytology*

Rabinowitz, D. and Betensky, R. A. (2000). Approximating the distribution of maximally selected McNemar's statistics. *Biometrics* **56**, 897-902.

Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics* **43**, 429-467.

Sauerbrei, W. (1999). The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics* **48**, 313-329.

Schulgen, G., Lausen, B., Olsen, J. and Schumacher, M. (1994). Outcome-oriented cutpoints in quantitative exposure. *American Journal of Epidemiology* **120**, 172-184.

Schumacher, M., Holländer, N., Sauerbrei, W. (1996). Reduction of bias caused by model building. In: American Statistical Association. *Proceedings of the Statistical Computing Section*, p. 1-7, American Statistical Association, Alexandria.

Schumacher, M., Holländer, N. and Sauerbrei, W. (1997). Resampling and cross-validation techniques: A tool to reduce bias caused by model building? *Statistics in Medicine* **16**, 2813-2827.

Van Houwelingen, H. C. and Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine* **9,** 1303-1325.

Verweij, P. and Van Houwelingen, H.C. (1993). Cross-validation in survival analysis. *Statistics in Medicine* **12**, 2305-2314.

Volinsky, C.T., Madigan, D., Raftery, A.E. and Kronmal, R.A. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society (Series C)* **46**, 433-448.

Worsley, K.J. (1982). An improved Bonferroni inequality and applications. *Biometrika* **69,** 297-302.