

# A SIMULATION STUDY OF CROSS-VALIDATION FOR SELECTING AN OPTIMAL CUTPOINT IN UNIVARIATE SURVIVAL ANALYSIS

DAVID FARAGGI AND RICHARD SIMON

*Biometric Research Branch, National Cancer Institute, 6130 Executive Blvd., Room 739, Rockville, MD 20852, U.S.A.*

## SUMMARY

Continuous measurements are often dichotomized for classification of subjects. This paper evaluates two procedures for determining a best cutpoint for a continuous prognostic factor with right censored outcome data. One procedure selects the cutpoint that minimizes the significance level of a logrank test with comparison of the two groups defined by the cutpoint. This procedure adjusts the significance level for maximal selection. The other procedure uses a cross-validation approach. The latter easily extends to accommodate multiple other prognostic factors. We compare the methods in terms of statistical power and bias in estimation of the true relative risk associated with the prognostic factor. Both procedures produce approximately the correct type I error rate. Use of a maximally selected cutpoint without adjustment of the significance level, however, results in a substantially elevated type I error rate. The cross-validation procedure unbiasedly estimated the relative risk under the null hypothesis while the procedure based on the maximally selected test resulted in an upward bias. When the relative risk for the two groups defined by the covariate and true changepoint was small, the cross-validation procedure provided greater power than the maximally selected test. The cross-validation based estimate of relative risk was unbiased while the procedure based on the maximally selected test produced a biased estimate. As the true relative risk increased, the power of the maximally selected test was about 10 per cent greater than the power obtained using cross-validation. The maximally selected test overestimated the relative risk by about 10 per cent. The cross-validation procedure produced at most 5 per cent underestimation of the true relative risk. Finally, we report the effect of dichotomizing a continuous non-linear relationship between covariate and risk. We compare using a linear proportional hazard model to using models based on optimally selected cutpoints. Our simulation study indicates that we can have a substantial loss of statistical power when we use cutpoint models in cases where there is a continuous relationship between covariate and risk.

## 1. INTRODUCTION

In many areas of medicine there is substantial interest in the identification of biological markers that can serve as prognostic or treatment selection factors. For example, a review of the search for prognostic indicators in patients with primary breast cancer is given by Gasparini *et al.*<sup>1</sup> Many markers are measured in laboratory assays as continuous variables. However, staging systems and clinical trial eligibility criteria generally require the expression of prognostic factors as categorical variables. For example, children with neuroblastoma who have amplification of the *n-myc* gene may be treated on a different protocol than other children.

A common practice is to choose a cutpoint that defines two risk groups for a continuously measured marker. The selection of a cutpoint often involves examination of different potential

cutpoints and choice of the one that minimizes the  $p$ -value associated with a comparison of outcome among patients with values above and below the cutpoint.

Several authors have pointed out that use of the same data to define the cutpoint and to evaluate statistical significance of the marker may produce distorted results. Hilsenbeck *et al.*<sup>2</sup> addressed this issue in relation to the question why the ‘significance’ of many prognostic markers is not confirmed in follow-up studies. They performed a simulation study to illustrate the high type I error rate that results from a simple search for the best cutpoint and subsequent declaration of the minimum  $p$ -value as significant if less than 5 per cent. They found that the type I error was about 25 per cent with 15 cutpoints examined and could reach 40 per cent when the number of cutpoints exceeded 50 for a two-sided test. They recommended use of a split sample approach. With the split sample approach one establishes the cutpoint with one part of the data and tests it with the other part. Silvestrini *et al.*<sup>3</sup> adopted this approach to evaluate p53 as an independent prognostic marker in breast cancer patients without axillary node involvement. The analysis entailed data from 256 patients, with 85 per cent of the data used to select the cutpoint and the remainder used for validation.

Lausen and Schumacher<sup>4</sup> provided a method for adjusting the  $\alpha$ -level of the logrank test<sup>5</sup> for the use of a maximally selected cutpoint. Their method applies to survival data and is an extension of the method of Miller and Sigmund<sup>6</sup> for the maximally selected chi-square statistic in  $2 \times 2$  tables. Altman *et al.*<sup>7</sup> provided a simple approximation to the Lausen and Schumacher correction. Although this approach corrects the type I error of the procedure, the estimation of the relative risk for the patients with marker values above and below the maximally defined cutoff remains biased.

To correct both the significance level and estimated relative risk, we propose a cross-validation approach<sup>8</sup> to select the cutpoint for continuous variables. The cross-validation procedure, in its simplest form, splits the data randomly into two halves, denoted by parts I and II. We select the cutpoint that minimizes the  $p$ -value on part I. Using this cutpoint, we assign all observations in part II to either group A (covariate above the cutpoint) or group B (covariate below the cutpoint). We then repeat the selection of a cutpoint on part II and use that cutpoint to assign all observations in part I to either group A or B. Once the procedure is completed each observation belongs to either group A or B. We then compute the two-sided stratified logrank test to determine if there is a significant difference between the risk groups. Complete details of the proposed method appear in Section 4. We conducted a simulation study using various cross-validation schemes to evaluate the size and power of the procedure. We compare the results from the cross-validation procedure with those obtained from the maximally selected logrank statistic with use of simulated data generated from a variety of changepoint models and under the null hypothesis. We provide the results for the simulation when the true changepoint is known, to serve as a baseline for the comparisons. We also performed a simulation with data generated from the Cox proportional hazards model<sup>9</sup> in which the covariate value influences hazard by a continuous logistic function. We compare the cross-validation cutpoint procedure to the adjusted  $\alpha$  level method and to the use of a proportional hazards model based on the assumption of a linear relationship between covariate value and log hazard.

## 2. EFFECT OF OPTIMALLY SELECTED CUTPOINT ON TYPE I ERROR

We began by evaluating the implications of naively using a set of data to both determine an optimally selected cutpoint for a covariate and for performing two sample significance tests comparing outcomes among groups defined based on that cutpoint. We generated 100 survival times from an independent and identically distributed (i.i.d.) exponential distribution,  $t_i \sim \exp(1)$

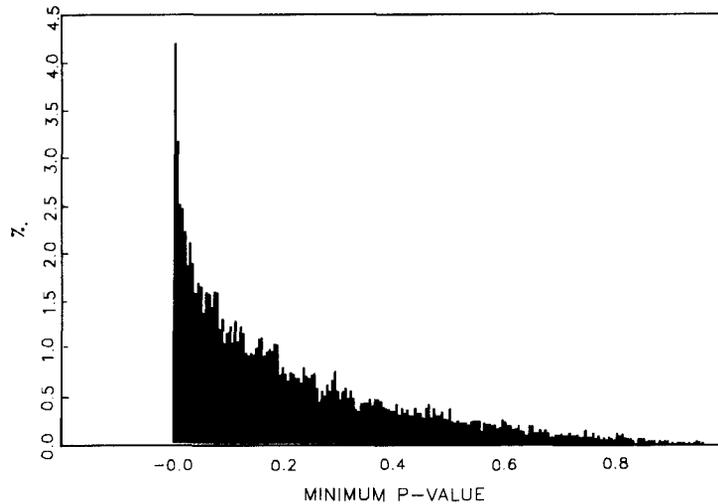


Figure 1. Minimal  $p$ -value distribution on 10,000 simulations

$i = 1, \dots, 100$ , and for each case a covariate  $x_i$  that is uniform  $[0, 1]$  and independent of  $t_i$ . For each simulated sample we performed an incremental search for the best cutpoint of  $x$ . Starting with a cutpoint value of 0.1 for the covariate, we grouped all observations that have the covariate value of less or equal to 0.1 into one risk group. Similarly we grouped all observations that have covariate value larger than 0.1 into the other risk group. Once we defined the risk groups, we performed the logrank test and calculated the two-sided  $p$ -value. We repeated this with an incremental increase of 0.05 in the definition of the cutpoint until we reached the value of 0.9 for the cutpoint. The best cutpoint was the one that achieved the minimum  $p$ -value. Figure 1 displays the histogram of the minimum  $p$ -values for 10,000 simulations. The width of each bar in the histogram is 0.005. Hence, for example, from the first bar we observe that the probability that the minimum  $p$ -value is in the range  $[0, 0.005]$  is 0.043. Since we had constructed this example under the null hypothesis that survival time and covariate value are independent, one might expect a uniform distribution of the  $p$ -values over  $[0, 1]$ . Figure 1 displays a strikingly non-uniform distribution of the minimum  $p$ -value. The probability that the minimum  $p$ -value is less than 0.05 is about 0.24.

### 3. MAXIMALLY SELECTED LOGRANK STATISTIC

Lausen and Schumacher,<sup>4</sup> following Miller and Sigmund,<sup>6</sup> derived the asymptotic null distribution of the maximally selected two-sided logrank test. Let  $[\varepsilon, 1 - \varepsilon]$   $0 < \varepsilon < 1$ , denote the range of quantiles of the prognostic factor values considered as cutpoints. Lausen and Schumacher restricted the search for the cutpoint to  $[\varepsilon, 1 - \varepsilon]$  rather than  $[0, 1]$  to utilize an asymptotic argument and to provide a reasonable amount of data in both groups. A correction to the logrank  $p$ -value is

$$P_{\text{corr}} = 4\phi(z)/z + \phi(z)[z - (1/z)]\log[(1 - \varepsilon)^2/\varepsilon^2] \quad (1)$$

where  $z$  is the  $(1 - P_{\text{min}}/2)$ -quantile of the standard normal distribution,  $P_{\text{min}}$  is the minimum logrank  $p$ -value obtained from the search and  $\phi$  denotes the standard normal density function.

Altman *et al.*<sup>7</sup> provided a simple approximation to (1). For example, for  $\varepsilon = 0.1$  the approximation is

$$P_{\text{corr}} \approx -1.63P_{\text{min}}(1 + 2.351 \log P_{\text{min}}). \quad (2)$$

They showed that this approximation is quite accurate for  $0.0001 \leq P_{\text{min}} \leq 0.1$ . Using approximation (2) we obtain that for  $P_{\text{corr}} \leq 0.05$  we require  $P_{\text{min}} \leq 0.002$ .

In addition to testing significance of the covariates on survival, one generally wishes to estimate the magnitude of the effect. This is usually accomplished by plotting the Kaplan–Meier curves for the subsets of patients with covariate values above and below the selected cutoff value and computing the relative risk for these two curves. As indicated by Altman *et al.*<sup>7</sup> although use of (2) will correct the size of the test, the Kaplan–Meier curves overestimate the true covariate effect on survival.

#### 4. CROSS-VALIDATION

One can use cross-validation to evaluate the significance of the optimally selected cutpoint and to estimate the relative risk. In general, a cross-validation procedure randomly splits the data into  $v$  sets of similar size. Leaving one set out, one estimates the parameters of interest using the remaining  $v - 1$  sets. One then uses the estimates to categorize or predict observations in the omitted set. For example, ten-fold cross-validation ( $v = 10$ ) randomly splits the data into ten parts of similar size. Ten times one uses 9/10 of the data for estimation and each time applies the estimates to the omitted 1/10th of the data.

To illustrate the procedure for cutpoint selection consider the two-fold cross-validation scheme:

1. Randomly split the data into two halves (denote the two subsets by I and II).
2. Estimate the cutpoint that minimizes the  $p$ -value for the two-sided logrank test using subset I.
3. Using the cutpoint obtained in step 2, assign each observation in subset II to either group A (covariate above cutpoint) or B (covariate below cutpoint).
4. Estimate the cutpoint that minimizes the  $p$ -value for the two-sided logrank test using subset II.
5. Using the cutpoint in step 4, assign each observation in subset I to either group A (covariate above cutpoint) or B (covariate below cutpoint).

Once the procedure is complete, all the observations in the sample have been assigned to either group A or B. We then compute the logrank statistic stratified by subset I and II. The key feature is that the cutpoint used to categorize each observation has been optimally selected from a subset that excludes the observation.

To estimate the relative risk, we used the Mantel–Haenszel hazard ratio estimator. Assume that  $x$  is a binary covariate with  $n_A$  patients having level A and  $n_B$  patients having level B. Denote the hazard rate at time  $t$ , for the patient with covariate level A, as  $h_A(t)$ . The proportional hazards model<sup>9</sup> assumes that

$$h_B(t) = \exp(\beta)h_A(t) \quad (3)$$

where  $h_B(t)$  is the hazard rate of patients with covariate at level B. Assume that there are  $k$  distinct times of death with no ties. We can write the  $k$  failure times as  $t_1 < t_2 < \dots < t_i < \dots < t_k$ , and

represent the data as a series of  $k$   $2 \times 2$  tables.<sup>5</sup> The  $i$ th table corresponding to  $t_i$  is

	Died	Survived	Total
Covariate at level A	$1 - x_i$	$n_{Ai} - 1 + x_i$	$n_{Ai}$
Covariate at level B	$x_i$	$n_{Bi} - x_i$	$n_{Bi}$
Total	1	$n_i - 1$	$n_i$

where  $x_i = 0$  if the patient that died had covariate level A, and  $x_i = 1$  if the level was B. The Mantel–Haenszel hazards ratio estimate<sup>10</sup> is

$$e^{\hat{\beta}_{MH}} = \frac{\sum_{i=1}^k x_i(n_{Ai} - 1 + x_i)/n_i}{\sum_{i=1}^k (1 - x_i)(n_{Bi} - x_i)/n_i}. \quad (4)$$

In Section 5 we use the reciprocal of the Mantel–Haenszel estimator to estimate the true relative risk in the simulation study.

Note that the cross-validation procedure is likely to provide different values of the ‘optimal’ cutpoint for subsets I and II. The procedure provides a categorization of each observation as either above or below the optimal cutpoint in the opposite subset. We calculate the stratified logrank statistic and the estimate of the relative risk for the two groups based on this binary categorization of all covariate values. The final cutpoint suggested for clinical use is, however, the one that optimizes the logrank statistic for the entire combined sample without cross-validation. Cross-validation serves to assess the significance of the cutpoint and to estimate the relative-risk associated with it. The fact that we might have used two cutpoints in the cross-validation has no relevance, however, for subsequent categorization of the covariate.

We investigated different  $v$ -fold cross-validation schemes in the simulation study. We investigated ten-fold, four-fold and two-fold schemes all with equal or approximately equal subsets of size  $200/v$ . We found that we obtained the best results in terms of the size of the test with use of two-fold cross-validation. Other cross-validation schemes produced type I error rates that exceeded the nominal value. For example, the ten-fold cross-validation procedure produced type I errors of about 11 per cent when the nominal value was 5 per cent. Thus, in Section 5 we report only the results for two-fold cross-validation.

## 5. SIMULATION STUDY

### 5.1. Dichotomized Covariate

We generated samples of 200 survival values using the Weibull survival distribution  $S(t) = e^{-t^{(1/\lambda)^\gamma}}$  and by varying the shape parameter  $\gamma = 1/3, 1/2, 1$  (exponential) and 2. We generated the covariates from the uniform  $[0, 1]$  distribution, independently of the survival times under the null hypothesis. Since varying the shape parameters of the Weibull distribution had little effect on the results, we report here only the results from the exponential distribution ( $\gamma = 1$ ). Under the alternative hypothesis we generated two risk groups. If the covariate value of an observation was smaller than the true prespecified change point, we assigned the observation to the low risk group. If the value of covariate corresponding to the observation was larger than the

changepoint, we assigned the observation to the high risk group. We set the scale parameter  $\lambda$  of the survival distribution as  $\lambda = 1$  for the high risk group. To generate observations for the low risk group, we chose  $\lambda$  so that the true relative risk ( $\text{TRR} = 1/\lambda^\gamma$ ) between the two groups was fixed at the same values for different choices of the shape parameter  $\gamma$ . We chose these TRR values to be 1.1, 1.2, 1.5 and 2.0. We also performed simulations that included random censoring with the procedure. Since the effect of censoring was to reduce the effective sample size with little change in the relative performance of the procedures investigated here, we report only the results without censoring.

We compared three procedures:

1. A procedure in which the true changepoint used to generate the survival values is also used for analysis (TR).
2. Optimal selection of the best cutpoint while correcting the  $p$ -value of the maximally selected logrank test with use of equation (2) (MS).
3. Two-fold cross-validation (CV).

All entries in the tables are averages over 1000 replications; in parenthesis we provide the standard errors of the averages. In all simulations we conducted a search for the best cutpoint over the range 0.10 to 0.90, using an increment of 0.05, on the uniformly  $[0, 1]$  distributed covariate. We calculated the logrank test only when both groups defined by the cutpoint had at least 15 observations. We applied this restriction because the logrank  $p$ -values are based on asymptotic normality.

As mentioned, the cross-validation procedure may result in estimation of two different cutpoints for the two halves of the data. In our simulations we calculated the mean of the absolute value of the difference between the two cutpoints and its standard error. These statistics were quite invariant across simulation experiments. For example, the mean absolute difference when the true changepoint was 0.2 (0.5) was 0.233 (0.205) with standard error of 0.006 (0.005). We emphasize that we provide the information on the difference in cutpoints for illustrative purposes only. We recommend the use of the cross-validation procedure only for testing the significance of the cutpoint and estimation of the hazard ratio. Once one has established the cutpoint as significant via the cross-validation procedure, one should use the full sample to establish a unique cutpoint for subsequent use. This cutpoint is selected in the following way. For each candidate cutpoint, all observations in the full sample are categorized as being above or below the cutpoint. One then computes a logrank  $p$ -value for this partition of the full sample into two groups. This process is repeated for all candidate cutpoints. The cutpoint which is associated with the smallest logrank  $p$ -value is selected to be recommended for future use.

When we generated the survival times independently of the covariate, all procedures provided type I error rates close to the correct nominal value of 5 per cent. The estimated relative risk using the simple search (MS) was positively biased by approximately 19 per cent with a standard error of approximately 2 per cent. The relative risk estimate obtained from cross-validation was approximately unbiased. The estimated bias was 1 per cent with a standard error of approximately 1 per cent.

Table I gives statistical power results for different true changepoints and relative risk values. For the maximally selected logrank statistic (MS) we used the adjusted  $p$ -values (equation (2)) to determine when to reject the null hypothesis.

The rows in Table I correspond to different values of the true relative risk (TRR) used to generate the data. For example, the first row in the table indicates TRR of 1.1. That is, for a preselected true changepoint, we evaluated the covariate value for each observation. If the covariate value was less than or equal to the true changepoint, we assigned the observation to the

Table I. Power for the Weibull distribution simulated data with  $\gamma = 1$  using different cutpoints and different true relative risk (TRR) values

TRR	True changepoint								
	0.2			0.3			0.5		
	TR	MS	CV	TR	MS	CV	TR	MS	CV
1.1	0.099 (0.021)	0.069 (0.013)	0.094 (0.020)	0.092 (0.020)	0.079 (0.014)	0.098 (0.021)	0.094 (0.021)	0.073 (0.014)	0.098 (0.021)
1.2	0.199 (0.028)	0.121 (0.018)	0.137 (0.024)	0.259 (0.031)	0.131 (0.020)	0.157 (0.026)	0.264 (0.032)	0.132 (0.020)	0.161 (0.026)
1.5	0.624 (0.034)	0.428 (0.032)	0.379 (0.034)	0.734 (0.031)	0.523 (0.035)	0.501 (0.035)	0.803 (0.028)	0.603 (0.035)	0.579 (0.035)
2.0	0.977 (0.011)	0.830 (0.028)	0.791 (0.033)	0.993 (0.006)	0.956 (0.018)	0.916 (0.020)	0.999 (0.002)	0.960 (0.014)	0.953 (0.015)

TRR denotes true relative risk. The TR (true) method uses a logrank test comparing outcome between the groups defined based on the true changepoint. The MS (maximally selected) method identifies the cutpoint that provides the maximum value of the logrank statistic and adjusts the significance level for the process of maximal selection. The CV (cross-validation) method determines a maximally selected cutpoint in a subsample and uses it to categorize observations in the other subsample. The table shows estimates of statistical power determined by computer simulation and the standard errors of the estimates

high risk group generated from the Weibull distribution with scale parameter  $\lambda = 1$ . If the covariate value was larger than the true changepoint, we assigned the observation to the low risk group and we chose  $\lambda$  so that the true relative risk between the two groups was 1.1.

Table I shows that for TRR = 1.1 the power obtained from cross-validation exceeds the power obtained using the maximally selected test and is close to the power obtained when we used the true changepoint that defined the risk groups. For TRR = 1.2 the power results obtained from the CV procedure are higher than those obtained from the MS procedure, although these differences are not as great as those obtained for TRR = 1.1. With both of these relative risk values, however, the power is quite low. For TRR = 1.5, the power achieved by the MS procedure is about 10 per cent larger than that obtained from the CV procedure. When the true relative risk between the two groups is 2.0, the power is reasonable for all of the methods, with some superiority (5 per cent) with use of MS. In almost all cases, the statistical power is considerably less for MS or CV than that which would be possible if the true value of the changepoint were known.

Table II shows the mean estimated relative risks and their standard errors for the three procedures. For TRR = 1.1 the estimated relative risks obtained from the CV procedure are close to true values while the MS procedure produces an upward bias in estimation of about 15 per cent. For TRR = 1.2 the estimated relative risks using CV are also close to the true values while the relative risks estimated by the MS procedure are biased upwards by about 10 per cent. For TRR = 1.5 the relative risks estimated by the MS procedure are again biased upwards by about 10 per cent while there is a downward bias of about 5 per cent with use of the CV procedure. For TRR = 2.0, the results are similar to those obtained for TRR = 1.5. The upward bias in estimation using the MS procedure is about 10 per cent and about twice as great as the downward bias in estimation using the CV procedure.

Table II. Estimated relative risk for the Weibull distribution simulated data with  $\gamma = 1$  using different cutpoints and different true relative risk (TRR) values

TRR	True changepoint								
	0.2			0.3			0.5		
	TR	MS	CV	TR	MS	CV	TR	MS	CV
1.1	1.13 (0.01)	1.26 (0.02)	1.09 (0.02)	1.09 (0.01)	1.29 (0.02)	1.10 (0.02)	1.11 (0.01)	1.28 (0.02)	1.11 (0.02)
1.2	1.23 (0.02)	1.36 (0.02)	1.17 (0.03)	1.23 (0.01)	1.36 (0.02)	1.18 (0.02)	1.22 (0.01)	1.34 (0.02)	1.19 (0.02)
1.5	1.55 (0.02)	1.62 (0.02)	1.43 (0.03)	1.52 (0.02)	1.63 (0.02)	1.45 (0.02)	1.51 (0.02)	1.63 (0.02)	1.45 (0.02)
2.0	2.07 (0.03)	2.12 (0.03)	1.92 (0.04)	2.04 (0.02)	2.12 (0.03)	1.95 (0.03)	2.03 (0.02)	2.11 (0.02)	1.93 (0.03)

TRR denotes true relative risk. The TR (true) method uses a logrank test comparing outcome between the groups defined based on the true changepoint. The MS (maximally selected) method identifies the cutpoint that provides the maximum value of the logrank statistic and adjusts the significance level for the process of maximal selection. The CV (cross-validation) method determines a maximally selected cutpoint in a subsample and uses it to categorize observations in the other subsample. The table shows estimates of relative risk determined by computer simulation and the standard errors of the estimates

## 5.2. Continuous Covariate

We also conducted a simulation experiment in which the true model had relative risk as a continuous function of the covariate value rather than a step function. We generated 200 survival values using the Weibull survival distribution  $S(t) = e^{-((1/\lambda)t)^\gamma}$ . We generated the covariates from the uniform  $[-1, 1]$  distribution, independently of the survival times under the null hypothesis. To generate the risks ( $\lambda_i$ ) we used the logistic function  $\lambda_i = e^{(c/[1 + \exp(-x_i\beta)])}$  with  $c = \ln(1.5)$  or  $c = \ln(2.0)$  so that when  $x_i\beta = -\infty$ ,  $\lambda_i = 1$  and when  $x_i\beta = +\infty$ ,  $\lambda_i = 1.5$  (2.0). Hence the relative risk at the extreme is 1.5 (2.0). We also varied  $\beta$  to obtain the relation between the value of the covariate and the risk from close to a linear relation ( $\beta = 1$ ) to a step function ( $\beta = \infty$ ). We performed simulations with shape parameter  $\gamma = 1/3, 1/2, 1$  and 2. Since varying this parameter had little influence, we report only results for the exponential distribution ( $\gamma = 1$ ).

Table III provides the statistical power estimated from the simulations that compared three models. The Cox proportional hazards model (PH) that uses a linear relation between the covariate and the risk, the maximally selected procedure (MS), and the cross-validation technique (CV) that dichotomizes the risk into two groups. The results are given for  $\beta = 1, 10$  and 20. We also provide in the table the standard error of the power estimates. From Table III we can see that when  $\beta = 1$ , that is, the relationship between the risk and the covariate is close to a linear relation, and  $c = \ln(1.5)$  there is 28 per cent loss of power (0.147 versus 0.106) when we compare the proportional hazards model with the cross-validation. The loss of power is even greater (55 per cent) when we compare the proportional hazards model with the maximally selected procedure (0.147 versus 0.066). For  $c = \ln(2.0)$ , we observe 28 per cent loss of power when we compare the proportional hazards model with the cross-validation results (0.264 versus 0.191), and almost 50 per cent loss of power when we compare the proportional hazards model with the maximally selected procedure (0.264 versus 0.133).

Table III. Power for the Weibull distribution simulated data with  $\gamma = 1$  using the logistic relation between the covariate and risk

$\beta$	$c = \ln(1.5)$			$c = \ln(2.0)$		
	PH	MS	CV	PH	MS	CV
1	0.147 (0.01)	0.066 (0.01)	0.106 (0.01)	0.264 (0.01)	0.133 (0.01)	0.191 (0.01)
10	0.631 (0.02)	0.383 (0.02)	0.445 (0.01)	0.972 (0.004)	0.903 (0.01)	0.900 (0.01)
20	0.655 (0.02)	0.402 (0.02)	0.492 (0.02)	0.984 (0.004)	0.906 (0.01)	0.904 (0.01)

$\beta$  denotes regression parameter in logistic function relating covariate to relative risk. The PH (proportional hazards) method uses a linear proportional hazards model relating survival to covariate value.  $c$  denotes the maximum value of the relative risk between individuals. The MS (maximally selected) method identifies the cutpoint that provides the maximum value of the logrank statistic and adjusts the significance level for the process of maximal selection. The CV (cross-validation) method determines a maximally selected cutpoint in a subsample and uses it to categorize observations in the other subsample. The table shows estimates of statistical power determined by computer simulation and the standard error of the estimates

As  $\beta$  increases, the relationship between the covariate  $x$  and the risk  $\lambda$  becomes more like a step function. That is, the natural logarithm of risk changes abruptly from 0 to  $c$  as the covariate  $x$  changes from negative to positive. For  $\beta = 20$  the power values obtained from all procedures increase; however, when  $c = \ln(1.5)$  there is still 25 per cent loss of power using the cross-validation procedure in comparison to the proportional hazards model (0.655 versus 0.492) and 39 per cent loss of power comparing the proportional hazards model with the maximally selected procedure (0.655 versus 0.402). For  $c = \ln(2.0)$  the loss of power is about 10 per cent comparing either the cross-validation or the maximally selected procedure with the proportional hazards model. These results point out that we can have a substantial loss of information when we use cutpoint models in cases where there is a continuous relationship between covariate value and relative risk. Even when we need cutpoints for practical reasons, it may be advisable to establish the significance of the covariate with use of a continuous model.

## 6. DISCUSSION

Many reports of medical research entail dichotomization of covariates as well as outcome variables. For example, variables such as oestrogen receptor status or oncogene expression of tumour cells, respiratory function,<sup>11</sup> blood glucose<sup>12</sup> and depression<sup>13</sup> are often dichotomized. In some cases, one views the biologic feature measured as essentially dichotomous, but the interpretation of assay results is complicated by non-homogeneous cell populations, contaminants and other sources of experimental variability. In other cases, we dichotomize essentially continuous measurements for ease of use. In either case, there has been little attention in the literature to the statistical issues involved in dichotomization of variables. For example, Ragland<sup>14</sup> studied the classification of patients as hypertensive based on their diastolic and systolic blood pressures. His recommendation was to investigate and report results for each possible cutpoint. He did not, however, address the statistical dangers of this approach.

We have evaluated two approaches to assess the consequences of an optimally selected cutpoint. It is not our purpose, however, to advocate the dichotomization of continuous prognostic factors. As shown in Table III, the use of a dichotomous covariate rather than a continuous model based on a linear approximation may result in a substantial loss of statistical power.

We have examined a cross-validation procedure for selecting a cutpoint for a continuous covariate. The procedure is in the spirit of the proposed split sample approach given by Hilsenbeck *et al.*<sup>2</sup> However, while the split sample approach uses one portion of the sample for the selection procedure with validation of the results on the other portion, the cross-validation procedure uses the complete sample and thus seems more efficient. We do not, however, provide numerical comparisons of the two methods in this paper. As noted before, the cross-validation procedure may provide different cutpoints for the two subsets of the data. One uses these two cutpoints, however, only at the stage of testing whether there are two different risk groups and in estimation of the relative risk. Once we reject the null hypothesis, we should use the full data set to determine a unique cutpoint for the covariate and it is that cutpoint that we recommend for future use. We see no purpose in even reporting the two cut-points obtained as part of the cross-validation.

We compared cross-validation with use of the maximally selected logrank procedure in terms of relative risk estimation and statistical power. Our simulation study has shown that the two-fold cross-validation procedure produced almost unbiased estimation of the relative risk when the true relative risk that generated the data was below 1.5. For relative risks above 1.5, the cross-validation procedure produced at most 5 per cent underestimation. The maximally selected logrank procedure overestimated the relative risk in all situations. The greatest bias existed when survival and covariate were independent. Both procedures produced approximately the correct  $\alpha$ -level under the null. For small values of relative risk, the cross-validation procedure achieved better power. However, when the relative risk increased, the power of the maximally selected logrank procedure was about 10 per cent higher than the power obtained from the cross-validation procedure. For large values of the relative risk, both procedures achieved high power.

As the sample size increases, one would expect that the degree of downward bias in estimation of relative risk for the cross-validation method would decrease. The bias exists because the maximally selected cutpoint in one sub-sample is only an estimate of the true breakpoint. When that estimate is used to categorize the patients in the other sub-sample, the resulting groups are each mixtures of subjects with covariate values both above and below the true breakpoint. The relative risk for these mixed groups is necessarily less than the relative risk between individuals above and below the true breakpoint.

We investigated different  $v$ -fold cross-validation schemes in the simulation study and obtained the best results in terms of the size of the test with use of two-fold cross-validation. We did not investigate using replicated  $v$ -fold cross-validation. Replicated two-fold cross-validation would involve repeating the two-fold cross-validation with different random splits into two subsamples. This would be quite computationally intensive because maximal cutpoint selection would have to be determined for each replication and the logrank statistic averaged over replications. Replicated cross-validation may, however, provide improved estimates of relative risk and increased statistical power and is worthy of further research.

Generally in survival data analysis, one models survival as a function of several covariates. If, for example, one wishes to dichotomize one covariate that corresponds to a new assay while incorporating the standard covariates in their original scales, the method proposed by Lausen and Schumacher<sup>4</sup> does not apply. It has been developed within the framework of the logrank statistic for testing a single binary covariate. On the other hand, one can easily generalize the

cross-validation procedure introduced here to accommodate this situation. By randomly splitting the sample, one selects the optimal cutpoint from one subset by minimizing the  $p$ -value associated with the parameter of the dichotomized covariate in, for example, the proportional hazards model instead of for a two-sample logrank test. One uses the optimal cutpoint from each subset to categorize the cases in the other subset. This suggested procedure with various correlations between the covariates is the subject for future research.

When dichotomization of a covariate is not an objective, there are a number of approaches one can use to relate covariate values to response. In addition to the usual methods of including linear or linear plus quadratic terms in a model such as Cox's proportional hazards model, there are available newer approaches based on smoothing splines,<sup>15</sup> regression splines<sup>16, 17</sup> and fractional polynomials.<sup>18</sup>

#### ACKNOWLEDGEMENT

The authors acknowledge the referees for helpful comments.

#### REFERENCES

1. Gasparini, G., Pozza, F. and Harris, A. L. 'Evaluating the potential usefulness of new prognostic and predictive indicators in node-negative breast cancer patients', *Journal of the National Cancer Institute*, **85**, 1206–1219 (1993).
2. Hilsenbeck, S. G., Clark, G. M. and McGuire, W. L. 'Why do so many prognostic factors fail to pan out', *Breast Cancer Research and Treatment*, **22**, 197–206 (1992).
3. Silvestrini, R., Benini, E., Daidone, M. G., Veneroni, S., Boracchi, P., Cappelletti, V., Di Fronzo, G. and Veronesi, U. 'P53 as an independent prognostic marker in lymph node-negative breast cancer patients', *Journal of the National Cancer Institute*, **85**, 965–970 (1993).
4. Lausen, B. and Schumacher, M. 'Maximally selected rank statistics', *Biometrics*, **48**, 73–85 (1992).
5. Mantel, N. 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemotherapy Reports*, **50**, 163–170 (1966).
6. Miller, R. and Sigmund, D. 'Maximally selected chi-square statistics', *Biometrics*, **38**, 1011–1016 (1982).
7. Altman, D. G., Lausen, B., Sauerbrei, W. and Schumacher, M. 'Dangers of using "optimal" cutpoints in the evaluation of prognostic factors', *Journal of the National Cancer Institute*, **86**, 829–835 (1994).
8. Efron, B. *The Jackknife the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia, 1982.
9. Cox, D. R. 'Regression models and life tables', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
10. Bernstein, L., Anderson, J. and Pike, M. C. 'Estimation of the proportional hazard in two-treatment-group clinical trials', *Biometrics*, **37**, 513–519 (1981).
11. Sedline, S., Tager, I., Speizer, F. E., Rosner, B. and Weiss, S. T. 'Longitudinal variability in airway responsiveness in a population-based sample of children and young adults: intrinsic and extrinsic contributing factors', *American Review Respiratory Disease*, **140**, 172–178 (1989).
12. Sugarman, J. and Percy, C. 'Prevalence of diabetes in Navajo Indian Community', *American Journal of Public Health*, **79**, 511–513 (1989).
13. Barnes, G. E., Currie, R. F. and Segall, A. 'Symptoms of depression in a Canadian urban sample', *Canadian Journal of Psychiatry*, **33**, 386–393 (1988).
14. Ragland, D. R. 'Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint', *Epidemiology*, **3**, 434–440 (1992).
15. Gray, R. J. 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association*, **87**, 942–951 (1992).
16. Harrell, F. E. Jr, Lee, K. L. and Pollock, B. G. 'Regression models in clinical studies: determining relationships between predictors and responses', *Journal of the National Cancer Institute*, **80**, 1198–1202 (1988).
17. Durrleman, S. and Simon, R. 'Flexible regression models with cubic splines', *Statistics in Medicine*, **8**, 551–561 (1989).
18. Royston, P. and Altman, D. G. 'Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling', *Applied Statistics*, **43**, 429–453 (1994).