Identifying a Time-dependent Covariate Effect in the Additive Risk Model

Cyrus Amir[1,2] and Ian W. McKeague[3,4]

[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23298-0032

[3]Department of Statistics, Florida State University, Tallahassee, FL 32306

SUMMARY

We develop a method to identify a time dependent covariate effect in the partly parametric additive risk model. The proposed method is based on a formal hypothesis test, whereas previously only an ad hoc procedure was available. Rates of convergence of restricted maximum likelihood estimators of regression coefficients based on the method of sieves play an important role in the development. We carry out a simulation study to assess the performance of the proposed test. We also apply our test to real data from a clinical trial on myelomatosis.

1 INTRODUCTION

In the Cox [1] proportional hazards model, the influence of the covariates is assumed to be constant over time except for variations in the baseline hazard function. This is an unrealistic assumption in many applications (for instance the prognostic value of

[2]Correspondence to: Cyrus Amir, Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23298-0032, email: camir@hsc.vcu.edu, Phone: (804) 827-2068, Fax: (804) 828-8900

[4]Contract/grant sponsor: NSF, contract/grant number: 9971784

a covariate measured at the start of a study may decline over the follow-up period), so it is important to develop methods of determining when time-dependent covariate effects are present. We study this problem within the framework of Aalen's [2] additive risk model. We are able to develop formal tests of hypothesis, where previously only ad hoc graphical methods were available. We show that our approach works well with simulated data and apply it to data from a clinical trial on myelamatosis.

The literature on Aalen's additive risk model includes papers of Aalen [3], Mau [4, 5], Henderson and Milner [6], Huffer and McKeague [7], and Henderson and Oman [8]. Corresponding to each covariate in Aalen's model, there is a time-varying function to be estimated, so, compared with Cox's model, relatively large samples are required to produce satisfactory results. To ameliorate this problem, McKeague and Sasieni [9] introduced the partly parametric additive risk model with conditional hazard function $\lambda(t|\boldsymbol{z}) = \boldsymbol{\alpha}'\boldsymbol{x} + \boldsymbol{\beta}(t)'\boldsymbol{y}$, where the covariate is partitioned as $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$. The covariate effects $\boldsymbol{\alpha}$ corresponding to $\boldsymbol{x}$ are constant, and those corresponding $\boldsymbol{y}$ are time dependent. McKeague and Sasieni constructed weighted least squares estimators of the regression coefficients. Use of these estimators, however, requires prior knowledge of the partition $(\boldsymbol{x}, \boldsymbol{y})$.

The problem of identifying time-dependent covariate effects in the additive risk model has been recognized since its inception. Aalen [2] analyzed survival times of lung cancer patients in a data set from the Veterans Administration Lung Cancer Study Group and noticed that one of the covariates had significant effect only over the first 100 day period of the study. Mau [4] recommended plots of the cumulative regression functions of Aalen's additive risk model, also known as Aalen plots, as a graphical means of detecting time-dependent covariate effects, and as a supplement to a Cox model analysis. Aalen [3] proposed to visually determine whether the effect of a covariate disappeared by looking at the pattern of bootstrapped estimates of the cumulative hazards. Henderson and Oman [8] considered an additive risk model analysis of the primary biliary cirrhosis of the liver data in Fleming and Harrington [10] (Appendix D). By inspecting graphs of the cumulative regression estimates, they concluded that two of the five cumulative hazard functions were not linear and hence that there were time-dependent covariate effects.

McKeague and Sasieni [9] in their study of the partly parametric additive risk model graphically identified covariates with time-dependent effects. They plotted 95% pointwise confidence intervals of the cumulative hazard of a covariate along with the corresponding straight line estimate of the cumulative hazard under the assumption that the covariate effect is constant. If some portion of the straight line fell outside the confidence intervals, the covariate was identified as having a time-dependent effect. They applied the partly parametric additive risk model to data from the British Medical Research Council's (1984) fourth myelomatosis trial and identified graphically that the covariate serum β_2 microglobulin had a time-varying effect.

Under a Cox regression model, Anderson and Senthiselvan [12], and Moreau, O'Quigley, and Mesbah [13] developed tests for identifying time-dependence of a covariate effect by dividing the time axis into segments (or bins) and then comparing the regression coefficients on these segments, see also Brown [11]. Murphy [14] studied asymptotic size and power of such a test when the number of bins increases with sample size. She showed that her test is asymptotically consistent against a wide class of alternatives.

In this article we adapt Murphy's [14] procedure to identify a time-dependent covariate effect in the partly parametric additive risk model. We divide the time axis into segments and then compute a normalized log-likelihood ratio statistic based on piecewise constant regression functions on these segments. In Section 4 we derive the asymptotic size of this test when the number of time segments increases appropriately with sample size. We apply the test to data from the British Medical Research Council's (1984) fourth myelomatosis trial to study whether the effect of serum β_2 microglobulin is time dependent. Derivation of the asymptotic size of the proposed test depends on rates of convergence of maximum likelihood sieve estimators of the regression coefficients; these rates are developed in Section 3. In Section 2, we specify the assumptions needed to carry out the proofs under an i.i.d. survival setup and introduce some notation. For brevity the proofs of the theorems have been omitted. They are available in detail in Technical Report 00-003 from Department of Biostatistics at Medical College of Virginia of Virginia Commonwealth University.

2 ASSUMPTIONS AND NOTATION

The observations are denoted $(\boldsymbol{Z}_i, T_i, \delta_i)$, $i = 1, \ldots, n$ and assumed to be i.i.d. Here $\boldsymbol{Z}_i$ is a bounded covariate vector, T_i a (possibly censored) failure time and δ_i a censoring indicator for the ith individual; $\delta_i = 1$ if T_i is uncensored. The censoring mechanism is assumed to be noninformative and conditionally independent of the underlying failure time, given the covariate. Let $N_i(t)$ denote the indicator of an uncensored failure prior to time t, and let $Y_i(t)$ denote a $\{0, 1\}$-valued process taking the value 1 when the ith individual is under observation. The conditional hazard of T_i is specified by the additive risk model $\lambda(t|\boldsymbol{Z}_i) = \boldsymbol{\beta}_0(t)'\boldsymbol{Z}_i$, where $\boldsymbol{\beta}_0(t)$ is a vector of regression functions having the same dimension as $\boldsymbol{Z}_i$, and $\lambda(t|\boldsymbol{Z}_i) > 0$ for all t.

Let $(I_1, \cdots, I_K)$ be indicators of $K = K_n$ consecutive time segments covering the follow-up period $[0, \tau]$, with the segment-lengths denoted $\boldsymbol{l} = (l_1, \cdots, l_K)$. Let $l_{(1)}$ and $l_{(K)}$ denote the minimum and maximum lengths, respectively. Define the matrix-valued function $\boldsymbol{v}(t) = E[\boldsymbol{Z}_i \boldsymbol{Z}_i' Y_i(t) / \lambda(t|\boldsymbol{Z}_i)]$.

We need the following assumptions:

A. The smallest eigenvalue of $\boldsymbol{v}(t)$ is bounded away from zero on $[0, \tau]$.

B. (B1) Each component of $\boldsymbol{\beta}_0(t)$ is Lipshitz; (B2) $\limsup_{n\to\infty} l_{(K)}/l_{(1)} < \infty$.

Under condition (B2), note that $|\boldsymbol{l}|^2 = O(l_{(K)}) = O(l_{(1)}) = O(K^{-1})$, where $|\cdot|$ is Euclidean norm. Let Θ_0 be the set of $(m+p)$-dimensional functions on $[0, \tau]$ with the first m components being time-varying functions and the last p components constant. Let Θ_K be the subset of Θ_0 in which the first m components are step functions having at most K jumps.

3 RESTRICTED MLE

In this section, we find rates of convergence for restricted maximum likelihood estimators of regression coefficients in the partly parametric additive risk model when some covariate effects are time dependent and the rest are constants. To derive the maximum likelihood estimators of the regression coefficients, we need to maximize the log-likelihood function,

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \int_0^\tau \ln(\boldsymbol{\beta}' \boldsymbol{Z}_i)\, dN_i - \int_0^\tau \boldsymbol{\beta}' \boldsymbol{Z}_i Y_i\, dt \right\}.$$

However, the space of all time-varying functions $\boldsymbol{\beta}(\cdot)$ is infinite dimensional and maximization of the log-likelihood $\mathcal{L}$ with respect to $\boldsymbol{\beta}$ does not lead to a valid estimator. A remedy under the method of sieves (Grenander [15]) is the selection of an increasing sequence of parameter spaces indexed by n, such that the log-likelihood can be maximized within each space, and whose union is dense in the original infinite dimensional space.

Friedman [16], McKeague [17], Leskow [18], and Murphy and Sen [19] have used the histogram sieve in the survival analysis context and Murphy [14] used the histogram sieve for hypotheses testing purposes. McKeague [17] studied sieve estimators of the regression coefficients in the Aalen's [2] additive risk model using the least squares method. Murphy and Sen [19] studied the restricted maximum likelihood estimators in a Cox-type regression model.

We maximize the log-likelihood over the *histogram sieve* Θ_K. Since we need to expand the log-likelihood about a point in Θ_K, preferably $\boldsymbol{\beta}_0$, and since $\boldsymbol{\beta}_0$ might not be in Θ_K, we need to find a point $\bar{\boldsymbol{\beta}}$ close to it which is in Θ_K. An arbitrary point makes the score function a martingale plus a bias term. We asymptotically eliminate this bias by choosing $\bar{\boldsymbol{\beta}}$ to be $\bar{\boldsymbol{\beta}}(\cdot) = \sum_{j=1}^{K} \bar{\boldsymbol{\beta}}_0^j I_j(\cdot)$, where the lth component of $\bar{\boldsymbol{\beta}}_0^j$ is

$$\bar{\beta}_0^{jl} = \frac{\int_0^\tau I_j(t)\beta_0^l(t)s_l(t)\,dt}{\int_0^\tau I_j(t)s_l(t)\,dt}, \qquad \text{for } j = 1, \cdots, K,$$

and $s_l(t) = E[Z_i^l Y_i(t)]$.

Since the log-likelihood is a concave function of $\boldsymbol{\beta}$, the restricted maximum likelihood estimators are unique. Our approach parallels Murphy and Sen [19] (Theorem 1) who studied sieved-partial maximum likelihood estimators in a Cox model setting. The following rate of convergence for the restricted maximum likelihood estimator was obtained by Taylor expanding the log-likelihood $\mathcal{L}(\boldsymbol{\beta})$ about $\bar{\boldsymbol{\beta}}$ and bounding its first, second and third partial derivatives at $\bar{\boldsymbol{\beta}}$.

Theorem 1 *Assume that Conditions* **A** *and* **B** *hold,* $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^1, \boldsymbol{\beta}_0^2)' \in \Theta_0$*, and* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\beta}}^2)'$ *maximizes* $\mathcal{L}(\boldsymbol{\beta})$ *over* Θ_K*. Also assume that* $n/K^2 \to \infty$ *and* $n/K^3 \to 0$*. Then*

$$\int_0^\tau \left|\widehat{\boldsymbol{\beta}}^1 - \boldsymbol{\beta}_0^1\right|^2 dt = O_{\mathcal{P}}\left((n|\boldsymbol{l}|^2)^{-1}\right) + O_{\mathcal{P}}\left(|\boldsymbol{l}|^4\right) \text{ and } n|\widehat{\boldsymbol{\beta}}^2 - \boldsymbol{\beta}_0^2|^2 = O_{\mathcal{P}}(1).$$

4 IDENTIFYING TIME DEPENDENCE

In this section we develop a standardized likelihood ratio test for identifying a time-dependent covariate effect in the partly parametric additive risk model and derive its asymptotic size. We apply this test to data from the British Medical Research Council's (1984) fourth myelomatosis trial and study the effect of serum β_2 microglobulin. We also conduct a simulation study to assess the performance of the test.

A simple test for identifying time-dependent covariate effects is to divide the time axis into a fixed number of intervals. Assume the covariate effects are constant over each interval and compare the regression coefficients of interest on these intervals. If their values differ greatly, we would conclude that the covariate effect of interest is time dependent. Since the number of intervals is fixed, however, this test is not consistent. For that reason we need to suitably increase in the number of bins with sample size.

Suppose we wish to identify whether a given component among the last p components of $\boldsymbol{\beta}_0$ is time-varying. That is, we wish to test the null hypothesis $H_0\colon \boldsymbol{\beta}_0 \in \Theta_0$ against the alternative H_1 that places no restriction on the given component. The corresponding standardized log-likelihood ratio based on sieves of the form Θ_K is written as

$$S_K = \frac{2\left[\mathcal{L}\Big(\widehat{\boldsymbol{\beta}}(H_{1K})\Big) - \mathcal{L}\Big(\widehat{\boldsymbol{\beta}}(H_{0K})\Big)\right] - (K-1)}{\sqrt{2(K-1)}} \tag{1}$$

where each $\widehat{\boldsymbol{\beta}}$ is the appropriately restricted maximum likelihood estimator defined in Section 3.

In the following theorem we show that the test statistic S_K converges in distribution to standard normal provided $K = K_n \to \infty$ at an appropriate rate. The proof parallels

Murphy [14] (Theorem 1). The idea is to add and subtract twice the log-likelihood at $\boldsymbol{\beta}_0$ to get

$$2\left[\mathcal{L}\left(\widehat{\boldsymbol{\beta}}(H_{1K})\right) - \mathcal{L}\left(\widehat{\boldsymbol{\beta}}(H_{0K})\right)\right] = 2\left[\mathcal{L}\left(\widehat{\boldsymbol{\beta}}(H_{1K})\right) - \mathcal{L}\left(\boldsymbol{\beta}_0\right)\right] - 2\left[\mathcal{L}\left(\widehat{\boldsymbol{\beta}}(H_{0K})\right) - \mathcal{L}\left(\boldsymbol{\beta}_0\right)\right].$$

Using a Taylor series expansion, we then rewrite each of the above terms as a quadratic form plus lower order terms. In the classical approach, the two quadratic forms combine to give the asymptotic normal distribution. However, we do not combine the quadratic forms. We standardize the quadratic forms and show that under H_0 the standardized version of the second term above converges to zero in probability and the standardized version of the first term is asymptotically normal. This is done by writing the quadratic form in terms of a martingale and applying a martingale central limit theorem.

Theorem 2 *Under the assumptions of Theorem 1, the test statistic S_K converges in distribution to standard normal.*

4.1 SIMULATION STUDY

To study the performance of the proposed test, we carried out a simulation study under various additive risk models. In each case, we tested for time-dependence in the effect of the second component z_2 for $n = 250$ and $n = 500$. The results are reported in Table I. The covariates were uniformly distributed on $[0, 1]$ and the censoring variables had exponential distributions with parameters adjusted to give the various censoring rates reported in the table. The tests were performed using $K = 5$ time segments, which was small enough to make at least one observation fall in each segment and large enough to make detection of time-dependence of the effect of z_2 feasible.

In the four additive risk models satisfying the null hypothesis, testing at the .05 nominal significance level we found observed significance levels of .05, .06, .06, and .06

for $n = 500$, and .05, .09, .09, and .06 for $n = 250$. Thus the test is slightly conservative at smaller sample sizes. The power for detecting H_1 in the model with two covariates and .10 significance level was .999 for $n = 500$ compared with .912 for $n = 250$. Our test has much higher power under the alternative with two covariates compared with three covariates.

4.2 APPLICATION

We applied the proposed test to data form the British Medical Research Council's (1984) fourth myelomatosis trial. Cuzick, Cooper and MacLennan [20] have indicated that the covariate serum β_2 microglobulin is of primary importance in predicting survival in myelomatosis patients. McKeague and Sasieni [9] fitted a partly parametric additive risk model to these data. They used Wald statistics to identify influential covariates and found serum β_2 microglobulin and haemoglobin to be influential. Using a graphical procedure, they concluded that effect of serum β_2 microglobulin was time dependent.

We tested for time-dependence in the effect of serum β_2 microglobulin in a model with serum β_2 microgobulin and haemoglobin as covariates, and a constant covariate to provide a baseline. Under the null hypothesis, serum β_2 microglobulin and haemoglobin had constant effects and the baseline hazard was time dependent. Under the alternative hypothesis, haemoglobin had constant effect and the baseline hazard and the effect of serum β_2 microglobulin were time-dependent. We chose 5 bins with equal width covering the time span of 0 to 2000 days of follow-up. The choice of $K = 5$ seems reasonable in view of the simulation results in the previous section; the sample size here is $n = 495$. In Table II, we report the parameter estimates under the null and alternative hypotheses. The standardized likelihood ratio test statistic was 3.819 (p-value 0.00013), providing strong evidence that the effect of serum β_2 microglobulin is time dependent.

5 CONCLUSION

We have developed a test statistic, (1), to identify a time-dependent covariate effect in the partly parametric additive risk model, and have derived its asymptotic null distribution. We have also derived the rate of convergence of restricted maximum likelihood estimators of the regression coefficients in the partly parametric additive risk model, which we used in the derivation of the asymptotic distribution of the sieve test statistic. We have conducted a simulation study to assess the performance of the test procedure, and have showed that it performs well for moderately large samples. We have then applied the test statistic to data from British Medical Research Council's (1984) fourth myelomatosis trial. We have found the effect of serum β_2 microglobulin to be time dependent and hence have confirmed the graphical results of McKeague and Sasieni [9].

AKNOWLEDGEMENTS

The research of the second author was supported by NSF grant 9971784.

REFERENCES

1. Cox DR. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* 1972; **34**, 187-220.

2. Aalen OO. A model for nonparametric regression analysis of counting processes. *Springer Lecture Series* 1980; 1-25.

3. Aalen OO. Further results on the non-parametric linear regression model in survival analysis. *Stat. Med.* 1993; **12**, 1569-1588.

4. Mau J. On a graphical method for the detection of time-dependent effects of covariates in survival data. *J. Roy. Statist. Soc. Ser. C* 1986; **35**, 245-255.

5. Mau J. A comparison of counting process models for complicated life histories. *Applied Stochastic Models and Data Analysis* 1988; **4**, 283-298.

6. Henderson R, Milner A. Aalen plots under proportional hazards. *J. Roy. Statist. Soc. Ser. C* 1991; **40**, 401-410.

7. Huffer FW, McKeague IW. Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Stat. Assoc.* 1991; **86**, 114-129.

8. Henderson R, Oman P. Influence in linear hazard models, *Scandinavian Journal of Statistics* 1993; **20**, 195-212.

9. McKeague IW, Sasieni PD. A partly parametric additive risk model. *Biometrika* 1994; **81**, 501-514.

10. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis.* Wiley: New York, 1991.

11. Brown CC. On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics* 1975; **31**, 863-872.

12. Anderson JA, Senthilselvan A. A two-step regression model for hazard functions. *Applied Statistics* 1982; **31**, 44-51.

13. Moreau T, O'Quigley J, Mesbah M. A global goodness-of-fit statistic for the proportional hazards model. *Applied Statistics* 1985; **34**, 212-218.

14. Murphy SA. Testing for a time dependent coefficient in Cox's regression model. *Scandanavian Journal of Statistics* 1993; **20**, 35-50.

15. Grenander U. *Abstract Inference.* Wiley: New York, 1981.

16. Friedman M. Piecewise exponential models for survival data with covariates *Annals of Statistics.* 1982; **10**, 101-113.

17. McKeague IW. A counting process approach to the regression analysis of grouped survival data. *Stochastic Processes. Appl.* 1988; **28**, 221-239.

18. Leskow J. Histogram maximum likelihood estimator of a periodic function in the multiplicative intensity model. *Statist. Decisions* 1988; **6**, 79-88.

19. Murphy SA, Sen PK. Time-dependent coefficients in a Cox-type regression model. *Stoch. Processes and their Appl.* 1991; **39** 153-180.

20. Cuzick J, Cooper EH, MacLennan IM. (1985). The prognostic value of serum β_2 microglobulin compared with other presentation features in myelomatosis. *British J. Cancer* 1985; **52**, 1-6.

Table 1: Observed size and power of the proposed test for identifying time-dependence of effect of z_2 in various additive risk models. Each observed size and power is based on 10,000 Monte Carlo samples.

Model $\lambda(t\|\boldsymbol{z})$	Censoring Rate	Bin Boundaries	Sample Size	Significance Level .01	.05	.10
$z_1 + z_2$	28.3%	$0, .25, .5, .75, 1, 5$	250	.0258	.0530	.0760
			500	.0239	.0512	.0724
$z_1 t + z_2$	37.6%	$0, .2, .4, .7, 1, 2$	250	.0531	.0914	.1217
			500	.0344	.0634	.0887
$t + z_2$	26.9%	$0, .2, .4, .7, 1, 2$	250	.0528	.0866	.1123
			500	.0303	.0582	.0838
$z_1 t + z_2 + z_3$	25.5%	$0, .17, .35, .65, 1, 2$	250	.0334	.0632	.0920
			500	.0286	.0576	.0782
$z_1 t + z_2 t$	37.4%	$0, .4, .69, .95, 1.25, 2.5$	250	.8221	.8857	.9121
			500	.9942	.9973	.9985
$z_1 t + z_2 t + z_3$	28.6%	$0, .2, .4, .7, 1, 2$	250	.3282	.4400	.5032
			500	.6145	.7252	.7768

Table 2: Restricted maximum likelihood estimates ($\times 10^4$) of the regression coefficients in the additive risk models for the British Medical Research Council's (1984) Fourth Myelomatosis Trial. Under H_0 the baseline hazard is time dependent, and the effects of haemoglobin and serum-β_2 are constant. Under H_1 the baseline hazard and the effect of serum-β_2 are time dependent and the effect of haemoglobin is constant. In the proposed test for time-dependence of the effect of serum-β_2, the above estimates result in a standardized likelihood ratio test statistic of 3.819. The p-value is .00013.

Hypothesis	Covariates	Bins				
		[0,400)	[400,800)	[800,1200)	[1200,1600)	[1600,2000]
H_0	baseline	10.54	10.48	13.03	18.60	10.56
	serum-β_2	4.999				
	haemoglobin	$-.05401$				
H_1	baseline	9.395	10.57	15.14	16.96	12.85
	serum-β_2	7.915	5.492	1.439	-2.749	1.022
	haemoglobin	$-.05674$				