

Testing for a Changepoint in the Cox Survival Regression Model

David M. Zucker¹, Sarit Agami², and Donna Spiegelman³

¹ Department of Statistics, Hebrew University, Mount Scopus, Jerusalem, Israel

Email: mszucker@mscc.huji.ac.il

² Department of Statistics, Hebrew University, Mount Scopus, Jerusalem, Israel

Email: sarit.agami@mail.huji.ac.il

³ Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston MA, USA

Email: stdls@hsph.harvard.edu

Abstract

The Cox regression model is a popular model for analyzing the effect of a covariate on a survival endpoint. The standard Cox model assumes that the covariate effect is constant across the entire covariate domain. However, in many epidemiological and other applications, there is interest in considering the possibility that the covariate of main interest is subject to a threshold effect: a change in the slope at a certain point within the covariate domain. In this paper, we discuss testing for a threshold effect in the case where the potential threshold value is unknown. We consider a maximum efficiency robust test (MERT) of linear combination form and supremum type tests. We present the relevant theory, present a simulation study comparing the power of various test statistics, and illustrate the use of the tests on data from the Nurses Health Study (NHS) concerning the relationship between chronic exposure to particulate matter of diameter 10 μm or less (PM_{10}) and fatal myocardial infarction. We also discuss power calculation for studies aimed at investigating the presence of a threshold effect, and present an illustrative power calculation. The simulation results suggest that the best overall choice of test statistic is a three-point supremum type test statistic. The power calculation methodology will be useful in study planning. Matlab software for performing the tests and power calculation is available by download from the first author's website.

AMS Classification: 62N03

Key words: Threshold, risk, score test, maximin efficiency robust test, supremum test

1 Introduction

The Cox (1972) model is a popular model for analyzing the effect of a covariate on a survival endpoint. The Cox model expresses the hazard function as

$$\lambda(t|\mathbf{z}(t)) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{z}(t)), \quad (1.1)$$

where $\lambda_0(t)$ is a baseline hazard function of unspecified form, $\mathbf{z}(t)$ is the covariate vector (which can depend on time), and $\boldsymbol{\beta}$ is vector of regression coefficients to be estimated. Estimation and inference theory for the Cox model is well established; see, for example, Kalbfleisch and Prentice (2002, Chs. 4-5) or Klein and Moeschberger (2003, Ch. 8). Model (1.1) assumes that the covariate effects are constant across the entire covariate domain. In many epidemiological and other applications, however, there is interest in considering the possibility that the covariate of main interest is subject to a threshold effect: a change in the slope at a certain point within the covariate domain.

Our interest in this issue was prompted by some instances of threshold effects observed in the Harvard-based Nurses' Health Study (NHS). For example, evidence of a threshold effect was observed in an analysis of data from the NHS concerning the relationship between exposure to particulate matter of diameter 10 μm or less (PM_{10}) and fatal myocardial infarction (Puett, 2009). We discuss these data in detail in Section 5. Evidence of a threshold effect was also observed in an analysis of data from NHS concerning the relationship between calcium intake and distal colon cancer (Wu et al., 2002). Among study participants followed over the period 1980-1996, with 164 distal colon cancer cases observed, no benefit of calcium intake was observed up to 700 mg/day, equivalent to two glasses of milk/day, while above that threshold a 30-40% benefit was observed.

Attention to nonlinear covariate effects, including threshold effects, plays an important role in setting guidelines for the maximum permissible exposure level of a given risk factor (Levy, 2003). Various studies of the health effects role of air pollution, for example, have revealed threshold effects that have implications for exposure level guidelines. In the American Cancer Society (ACS) analysis of chronic $\text{PM}_{2.5}$ exposure from air pollution in relation to lung cancer mortality (Pope et al., 1995), a steep rise in risk is seen up to approximately 15 g/m^3 and then the dose-response curve appears to flatten out. A re-analysis of the ACS data estimated a stronger relationship of $\text{PM}_{2.5}$ with overall mortality up to 16 g/m^3 , with statistically significant evidence for non-linearity (Abrahamowicz et al., 2003). The evidence for thresholds in acute air pollution effects has been studied further by other researchers (Samoli et al., 2005; Daniels et al., 2000; Smith et al., 2000).

Most of these studies did not find statistically significant evidence of departure from linearity, but the lack of statistical significance could be due to low statistical power. Roberts and Martin (2006) discussed the statistical aspects of assessing nonlinearities in the relationship between aerial particulate matter concentration and mortality. They noted that many studies used the Akaike information criterion (AIC) (Akaike, 1974) as a model selection technique, and they presented simulation results showing that the AIC approach performs inadequately in detecting nonlinearities when they exist. When nonlinear effects are hypothesized, epidemiologists often categorize the continuous exposure variable, accepting the well known loss of power with this approach (Greenland, 1995). A better strategy for assessing threshold effects is called for. The aim of this paper is to discuss better strategies.

Let $X(t)$ denote the covariate of main interest and $\mathbf{W}(t) \in \mathbb{R}^p$ the vector of additional covariates. We then have the model

$$\lambda(t|x(t), \mathbf{w}(t)) = \lambda_0(t) \exp(\boldsymbol{\gamma}^T \mathbf{w}(t) + \beta x(t) + \omega(x(t) - \tau)_+), \quad (1.2)$$

for some changepoint τ , where $u_+ = \max(u, 0)$. This type of changepoint model has been extensively investigated in the case of classical linear regression (Seber and Wild, 1989, Chapter 9). Küchenhoff and Wellich (1997) have examined such models in the setting of the generalized linear model.

Two basic cases can be identified: the case where the prospective changepoint τ is known and the case where τ is unknown. When τ is known, inference for ω can be carried out in a standard manner in the framework of the model (1.1), with $(z_1(t), \dots, z_p(t)) = \mathbf{w}(t)$, $z_{p+1}(t) = x(t)$, and $z_{p+2}(t) = (x(t) - \tau)_+$. The Cox partial likelihood estimation procedure and inference for the parameters can be carried out in the usual way. The case where τ is unknown, which is the case of greater relevance in applications, is more complex. A difficulty in estimation is posed by the fact that the partial likelihood is discontinuous in τ , so that the standard estimation approach based on differentiating the likelihood function does not work. The typical strategy for estimation is to estimate $\boldsymbol{\gamma}$, β , ω for each given τ using the standard partial likelihood method for the Cox model, and then

perform a grid search across the covariate domain to identify the value of τ at which the partial likelihood is maximized. The analogous estimation strategy in the case of linear regression has been examined in depth from a theoretical standpoint by Feder (1975). In a recent paper, Kosorok and Song (2007) carried out a theoretical analysis of changepoint inference for transformation survival models, which covers the Cox model as a special case.

Here, we are interested in testing for the existence of a changepoint, i.e., in testing the null hypothesis $H_0 : \omega = 0$. This is a nonstandard hypothesis testing problem, since the changepoint parameter (τ) is undefined under the null hypothesis. Consequently, standard asymptotic theory does not apply. Davies (1977, 1987) provides a general discussion of hypothesis testing when the nuisance parameter is undefined under the null hypothesis. Zucker and Yang (2005) discuss a problem of this sort in the context of a family of survival models. There are two main approaches in the literature for such problems. The first is Gastwirth's (1966, 1985) maximin efficient robust testing (MERT) approach; the second is the supremum (SUP) test approach of Davies (1977, 1987). In the case of changepoint inference with survival data, Kosorok and Song (2007) took the SUP approach.

The purpose of this paper is to provide a basic overview of the options for testing for a covariate threshold in the Cox model, compare the power of the various tests, and present methods for power and sample size calculations for this testing problem. Section 2 presents the setting and notation, and provides some background theory. Section 3 describes the various test statistics considered. Section 4 presents a simulation study to assess finite sample Type I error and statistical power. The range of scenarios we consider is broader than in the limited simulation study in Kosorok and Song's (2007) paper, which was primarily devoted to theory: We consider several values of the sample size, true changepoint, and initial slope β . In addition, along with the "full" supremum test, we consider the two-point and three-point supremum tests SUP2 and SUP3 described by Zheng and Chen (2005). These two tests, which are much easier to implement than the full supremum test, were not considered by Kosorok and Song. Section 5 provides an illustration of the tests on the above-mentioned data from the NHS on the relationship between PM₁₀ exposure and risk of fatal myocardial infarction. Section 6 discusses power calculation, providing the necessary theory and an illustrative example. Section 7 provides a brief discussion.

2 Setting, Notation, and Background

We work under a standard survival analysis setup. We have observations on n independent individuals. For a given individual i , $(\mathbf{W}_i(t), X_i(t))$ denotes the covariate process, T_i° denotes the survival time, and C_i denotes the time of right censoring. Again, X is the covariate of main interest, which is subject to a possible threshold effect, while \mathbf{W} is a vector of additional covariates. The observed data consist of the observed follow-up time $T_i = \min(T_i^\circ, C_i)$, the event indicator $\delta_i = I(T_i^\circ \leq C_i)$, and the covariate process $(\mathbf{W}_i(t), X_i(t))$ for $t \in [0, T_i]$. We allow left-truncation, such as in studies where the time metameter is age and people enter the study at different ages. With \tilde{T}_i denoting the time of entry to the study, we let $Y_i(t) = I(\tilde{T}_i \leq t \leq T_i)$ denote the at-risk indicator. We denote the maximum possible follow-up time by t_{max} . We assume that, conditional on the covariate process, the censoring is noninformative for the survival time in the sense of Andersen et al. (1993, Sec. III.2.3). Finally, we assume that the survival time T_i° follows the model (1.2). We write $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \beta, \omega, \tau)^T$ and $\boldsymbol{\theta}' = (\boldsymbol{\gamma}^T, \beta)^T$. We let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\gamma}}^T, \tilde{\beta}, \tilde{\omega}, \tilde{\tau})^T$ denote the true value of $\boldsymbol{\theta}$.

We now present some definitions and theoretical results which we will use in the next section. We define

$$S_0(t, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n Y_i(t) e^{\boldsymbol{\gamma}^T \mathbf{W}_i(t) + \beta X_i(t) + \omega(X_i(t) - \tau)_+}, \quad s_0(t, \boldsymbol{\theta}) = E_0[Y_i(t) e^{\boldsymbol{\gamma}^T \mathbf{W}_i(t) + \beta X_i(t) + \omega(X_i(t) - \tau)_+}],$$

where E_0 denotes expectation under the null hypothesis. Further, for given functions g, g_1 , and g_2 , we define

$$\begin{aligned} S(t, \boldsymbol{\theta}, g) &= \frac{1}{n} \sum_{i=1}^n Y_i(t) g(\mathbf{W}_i(t), X_i(t)) e^{\boldsymbol{\gamma}^T \mathbf{W}_i(t) + \beta X_i(t) + \omega(X_i(t) - \tau)_+}, \\ s(t, \boldsymbol{\theta}, g) &= E_0[Y_i(t) g(\mathbf{W}_i(t), X_i(t)) e^{\boldsymbol{\gamma}^T \mathbf{W}_i(t) + \beta X_i(t) + \omega(X_i(t) - \tau)_+}], \\ S(t, \boldsymbol{\theta}, g_1, g_2) &= \frac{1}{n} \sum_{i=1}^n Y_i(t) g_1(\mathbf{W}_i(t), X_i(t)) g_2(\mathbf{W}_i(t), X_i(t)) e^{\boldsymbol{\gamma}^T \mathbf{W}_i(t) + \beta X_i(t) + \omega(X_i(t) - \tau)_+}, \\ s(t, \boldsymbol{\theta}, g_1, g_2) &= E_0[Y_i(t) g_1(\mathbf{W}_i(t), X_i(t)) g_2(\mathbf{W}_i(t), X_i(t)) e^{\boldsymbol{\gamma}^T \mathbf{W}_i(t) + \beta X_i(t) + \omega(X_i(t) - \tau)_+}]. \end{aligned}$$

Next, for a given function g , we define

$$U(g, \boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \left[g(\mathbf{W}_i(t), X_i(t)) - \frac{S(T_i, \boldsymbol{\theta}, g)}{S_0(T_i, \boldsymbol{\theta})} \right].$$

Under mild regularity conditions, the following results can be shown using established counting process theory (Fleming and Harrington, 1991). First, $U(g, \bar{\boldsymbol{\theta}})$ is asymptotically mean-zero normal with variance

$$V(g) = \int_0^{t_{max}} \left[\frac{s(t, \bar{\boldsymbol{\theta}}, g, g)}{s_0(t, \bar{\boldsymbol{\theta}})} - \left(\frac{s(t, \bar{\boldsymbol{\theta}}, g)}{s_0(t, \bar{\boldsymbol{\theta}})} \right)^2 \right] s_0(t, \bar{\boldsymbol{\theta}}) \lambda_0(t) dt. \quad (2.1)$$

Second, for two functions g_1 and g_2 , the random vector $[U(g_1, \bar{\boldsymbol{\theta}}), U(g_2, \bar{\boldsymbol{\theta}})]$ is asymptotically mean-zero bivariate normal, with the covariance $C(g_1, g_2) = \text{Cov}(U(g_1, \bar{\boldsymbol{\theta}}), U(g_2, \bar{\boldsymbol{\theta}}))$ given by

$$C(g_1, g_2) = \int_0^{t_{max}} \left[\frac{s(t, \bar{\boldsymbol{\theta}}, g_1, g_2)}{s_0(t, \bar{\boldsymbol{\theta}})} - \left(\frac{s(t, \bar{\boldsymbol{\theta}}, g_1)}{s_0(t, \bar{\boldsymbol{\theta}})} \right) \left(\frac{s(t, \bar{\boldsymbol{\theta}}, g_2)}{s_0(t, \bar{\boldsymbol{\theta}})} \right) \right] s_0(t, \bar{\boldsymbol{\theta}}) \lambda_0(t) dt. \quad (2.2)$$

Finally, under the null hypothesis $H_0 : \omega = 0$, the covariance $C(g_1, g_2)$ can be consistently estimated by

$$\hat{C}(g_1, g_2) = \frac{1}{n} \sum_{i=1}^n \delta_i \left[\frac{S(T_i, \hat{\boldsymbol{\theta}}'_0, *, g_1, g_2)}{S_0(T_i, \hat{\boldsymbol{\theta}}'_0, *)} - \left(\frac{S(T_i, \hat{\boldsymbol{\theta}}'_0, *, g_1)}{S_0(T_i, \hat{\boldsymbol{\theta}}'_0, *)} \right) \left(\frac{S(T_i, \hat{\boldsymbol{\theta}}'_0, *, g_2)}{S_0(T_i, \hat{\boldsymbol{\theta}}'_0, *)} \right) \right], \quad (2.3)$$

and the variance $V(g)$ by $\hat{V}(g) = \hat{C}(g, g)$, where $\hat{\boldsymbol{\theta}}'_0$ denotes the partial likelihood estimate of $\boldsymbol{\theta}'$ for the null hypothesis model, and where, in the notation $\boldsymbol{\theta} = (\boldsymbol{\theta}', *)$, the asterisk denotes evaluation under $\omega = 0$ (in which case τ is immaterial).

3 Test Statistics

3.1 Known changepoint

When the prospective changepoint τ is known, we can test $H_0 : \omega = 0$ in a standard way. In this context, the parameter ω is the parameter of main interest, while $\boldsymbol{\theta}' = (\boldsymbol{\gamma}^T, \beta)^T$ is a nuisance parameter. Define $g_j(\mathbf{w}, x) = w_j$ for $j = 1, \dots, p$, $g_{p+1}(x) = x$, and $g_{p+2}(x) = (x - \tau)_+$. The elements of the Cox partial likelihood score vector are given by $U_j(\boldsymbol{\theta}) = U(\boldsymbol{\theta}, g_j)$. The partial likelihood score statistic for testing H_0 is given by

$$W_\tau = \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \left[g_{p+2}(\mathbf{W}_i(t), X_i(t)) - \frac{S_1(T_i, \hat{\boldsymbol{\theta}}'_0, *, g_2)}{S_0(t, \hat{\boldsymbol{\theta}}'_0, *)} \right]. \quad (3.1)$$

Using the theory presented in the preceding section, we find that the null hypothesis distribution of $\mathbf{U}(\bar{\boldsymbol{\theta}})$ is mean-zero multivariate normal, with covariance matrix $[C(g_j, g_k)]_{j,k=1, \dots, p+2}$. Now, let \mathbf{C} denote the matrix $[C(g_j, g_k)]_{j,k=1, \dots, p+1}$, and let $\mathbf{d}(\tau)$ denote the column vector with components $C(g_j, g_{p+2}), j = 1, \dots, p+1$. Then, from with standard theory for score tests with estimated nuisance parameters (Cox and Hinkley, 1974, pp. 321-324), the asymptotic null distribution of W_τ is mean-zero normal with variance

$$V_\tau = C(g_{p+2}, g_{p+2}) - \mathbf{d}(\tau)^T \mathbf{C}^{-1} \mathbf{d}(\tau)$$

This variance can be consistently estimated using (2.3), and we denote the estimate by \hat{V}_τ . We define $W_\tau^* = W_\tau / \sqrt{\hat{V}_\tau}$, which has a standard normal asymptotic null distribution.

In some cases, investigators may specify an a priori guess of the changepoint, and carry out a test assuming that the true changepoint is known to be equal to the guessed value. In Section 4, we will discuss the power loss incurred by guessing wrong.

3.2 MERT statistic

In this and the next sub-section, we assume that the true τ (if there is a change at all) is posited to lie in some prespecified range $[\tau_{min}, \tau_{max}]$. Unless explicitly stated otherwise, all τ values referred to from now on will be in

this range. In the simulation study presented in Section 4, we will examine the power of the tests under various values of the true τ , both within and outside the posited range.

Let us consider two possible τ values τ_1 and τ_2 , and suppose that the true τ value is τ_2 . Then, as in Gastwirth (1966, 1985), the Pitman asymptotic relative efficiency (ARE) of $W_{\tau_1}^*$ relative to $W_{\tau_2}^*$ is equal to the square of the asymptotic null correlation $\rho(W_{\tau_1}^*, W_{\tau_2}^*)$ between $W_{\tau_1}^*$ and $W_{\tau_2}^*$. Using the results presented in Section 2, we find that

$$\rho(\tau_1, \tau_2) \equiv \rho(W_{\tau_1}^*, W_{\tau_2}^*) = \frac{\kappa(\tau_1, \tau_2)}{\sqrt{V_{\tau_1} V_{\tau_2}}}, \quad (3.2)$$

where, setting $h_1(x) = (x - \tau_1)_+$ and $h_2(x) = (x - \tau_2)_+$, we have

$$\kappa(\tau_1, \tau_2) = C(h_1, h_2) - \mathbf{d}(\tau_1)^T \mathbf{C}^{-1} \mathbf{d}(\tau_2).$$

For a given set of τ values τ_1, \dots, τ_K and positive coefficients a_1, \dots, a_K , and a specified true τ value τ^* , the ARE of the linear combination statistic

$$Q(\mathbf{a}) = \sum_{k=1}^K a_k W_{\tau_k}^* \quad (3.3)$$

relative to $W_{\tau^*}^*$ is equal to the square of the asymptotic null correlation $\rho(Q(\mathbf{a}), W_{\tau^*}^*)$ between $Q(\mathbf{a})$ and $W_{\tau^*}^*$, which is given by

$$\rho(Q(\mathbf{a}), W_{\tau^*}^*) = \frac{\sum_{k=1}^K a_k \rho(\tau_k, \tau^*)}{\sum_{k=1}^K \sum_{l=1}^K a_k a_l \rho(\tau_k, \tau_l)}.$$

The Maximin Efficiency Robust Test (MERT) statistic is defined to be the linear combination statistic $Q(\mathbf{a})$ for which the worst-case ARE, namely, $\min_{\tau^* \in \mathcal{T}} \rho(Q(\mathbf{a}), W_{\tau^*}^*)$, is as high as possible, where \mathcal{T} denotes the set of τ values under consideration. When \mathcal{T} consists of a finite set of points, it is possible, as described in Gastwirth (1966), to find the optimal linear combination using quadratic programming. In our case, the set \mathcal{T} consists of the entire continuous range $[\tau_{min}, \tau_{max}]$, but the MERT over a moderately fine finite grid will be a reasonable approximation to the MERT over the entire continuous range.

3.3 Supremum statistics

The supremum statistic is defined as $\text{SUP} = \sup_{\tau \in [\tau_{min}, \tau_{max}]} |W_{\tau}^*|$. Davies (1977, 1987), working in a general setting, presented some approximations to the distribution of this statistic. In the setting of our problem, Kosorok and Song (2007) also considered this statistic, and used a weighted bootstrap scheme to obtain critical values.

In practice, it is possible to replace the supremum over the entire continuous range $[\tau_{min}, \tau_{max}]$ by the supremum over a moderately fine grid. With this implementation, the critical value can be computed using established routines for computing multivariate normal probabilities (Genz and Bretz, 1999, available in Matlab, R, and other packages).

Alternative statistics in the same spirit are the SUP2 and SUP3 statistics considered by Zheng and Chen (2005). These statistics are defined as

$$\text{SUP2} = \max(|W_{\tau_{min}}^*|, |W_{\tau_{max}}^*|), \quad \text{SUP3} = \max(|W_{\tau_{min}}^*|, |W_{\tau_{mid}}^*|, |W_{\tau_{max}}^*|),$$

where τ_{mid} is some intermediate value between τ_{min} and τ_{max} , such as the midpoint.

4 Simulation Study

In this section, we present a simulation study comparing the MERT, SUP, SUP2, and SUP3 statistics. For the SUP statistic, we used Kosorok and Song's (2007) weighted bootstrap scheme to compute critical values. We present the power of the various statistics for several scenarios. As a benchmark, we also present the power of the optimal score statistic if the true changepoint were known.

We performed the simulations under a setup with a single time-independent covariate X , without additional background covariates \mathbf{W} . We worked under fixed administrative censoring at time $t_{max} = 10$. We considered two configurations of the parameters β and ω : (a) $\beta = 0$ and $\omega = \log(1.5)$, corresponding to a log hazard which is flat over the range of X up to the changepoint and then slopes upward; (b) $\beta = \log(1.2)$ and $\omega = \log(1.5)$,

corresponding to a log hazard with a modest initial slope and a slightly steeper slope after the changepoint. We took $\lambda_0(t)$ equal to a constant value λ_0 , corresponding to exponential survival, with λ_0 chosen so as to yield 75% censoring under the relevant value for β and $\omega = 0$. The covariate X was generated as standard normal. We considered two possible sets for the true changepoint value, each defined by a five-point grid τ_1, \dots, τ_5 . Grid A consisted of the values $\Phi^{-1}(0.1), \Phi^{-1}(0.25), \Phi^{-1}(0.5), \Phi^{-1}(0.75), \Phi^{-1}(0.9)$. Grid B consisted of the values $\Phi^{-1}(0.3), \Phi^{-1}(0.4), \Phi^{-1}(0.5), \Phi^{-1}(0.6), \Phi^{-1}(0.7)$. Grid A, the wider one, is more realistic for the epidemiological applications that motivated our work.

The implementation of the MERT statistic was based on these grids. The SUP2 statistic was computed as $\text{SUP2} = \max(|W_{\tau_1}^*|, |W_{\tau_5}^*|)$, while the SUP3 statistic was computed as $\text{SUP3} = \max(|W_{\tau_1}^*|, |W_{\tau_3}^*|, |W_{\tau_5}^*|)$. The sample sizes considered were $n = 500, 1000, 2000$. Power was computed based on two-sided testing at the 0.05 level. For each of the two grids, we also computed power for the case where the true τ value was outside the specified grid. For Grid A, the outside τ values were $\Phi^{-1}(0.05)$ and $\Phi^{-1}(0.95)$; for Grid B, the outside τ values were $\Phi^{-1}(0.2)$ and $\Phi^{-1}(0.8)$. We ran 1000 simulation replications for each case.

Table 1 presents critical values for the various tests based on asymptotic normal theory, and corresponding Type I error levels. Tables 2 and 3 show the power results. In generating these tables, we used normal-theory critical values for all tests except for SUP, for which we used Kosorok and Song's (2007) weighted bootstrap scheme. The version of SUP with normal-theory critical values gave similar results. We also computed power for the score test assuming various *a priori* guessed values of the changepoint, at different distances from the true changepoint. We do not show these results in the tables, but we will describe the findings.

The patterns that emerged were as follows. For all of the tests, including the optimal score test, the power is greater when the true changepoint is near the center of the covariate domain than when it is in the extremes. This finding is expected, because when the changepoint is in one of the extremes, we have little data on one side of the changepoint. Regarding the power of the score test with a guessed value for the changepoint, we found, as expected, that the power is relatively high when the guessed changepoint is close to the true changepoint, and relatively low when the guessed changepoint is far off the mark. In the latter case, the MERT and SUP tests performed much better. In regard to the comparison between MERT and SUP, different results were seen for the two grids considered. For Grid A, the one with the wider range, SUP generally showed higher power than MERT, whereas for Grid B, the opposite was seen. Both patterns became more pronounced with increasing sample size.

Additional insight into these results can be gained by considering the null correlation between the pair of score statistics corresponding to the two τ values at the extremes of the range considered, which, as noted above, expresses the ARE of a given member of the pair relative to the other member when the other member is optimal. The null correlation for this extreme pair is about 0.25 for Grid A and about 0.70 for Grid B. Thus, when the null correlation of the extreme pair was low, SUP was favored over MERT, whereas when this correlation was high, MERT was favored over SUP. This finding was consistent with Freidlin, Podgor, and Gastwirth's (1999) power comparisons of MERT versus SUP in other settings. In regard to the comparison of SUP, SUP2, and SUP3, we found that (a) SUP and SUP3 are generally similar, with neither coming out clearly better than the other, (b) as between SUP or SUP3 and SUP2, the SUP2 statistic was better, unsurprisingly, when the true changepoint was near one of the extremes, but it was worse, sometimes substantially, when the true changepoint was in the middle of the range. In comparison with the optimal score test assuming the true changepoint is known, the MERT and SUP tests performed well under Grid B, the one with the narrow range, whereas, under Grid A, substantially lower power is seen. When the true τ value is outside the specified range, there is a substantial drop in the power with the MERT and SUP test compared with the optimal test when the changepoint is known, and the SUP approach generally yielded higher power than the MERT approach.

5 Illustrative Example

As noted in the introduction, our work was motivated by some instances of threshold effects observed in the Harvard-based Nurses' Health Study (NHS), including threshold effects observed in the NHS's investigation of the long term health effects of air pollution. We consider here an analysis of NHS data concerning the effect of exposure to particulate matter of diameter 10 μm or less (PM_{10}) in relation to incidence of fatal myocardial infarction (MI) (Puett et al., 2009). Here, 92,993 female nurses were followed from June 1992 to June 2006, with 1,050 fatal MI events observed. PM_{10} exposure was assessed for each individual on a monthly basis by linking the individual's residential address to her predicted PM_{10} exposure using a spatio-temporal model derived from

data from EPA area monitors (Yanosky et al., 2008; Paciorek et al., 2009). The monthly predicted exposures were summed to form a 12-month PM₁₀ moving average PM10MA, which constituted the exposure variable of interest. PM10MA is a time-varying covariate. In addition to the main covariate PM10MA, the analysis adjusted for confounding by the following covariates (all time-varying): calendar year, indicator variables for season, and indicator variables for US state of residence. The time scale in the analysis was age in months, so that the data are subject to left-truncation, which was handled by appropriate definition of the at-risk indicator function, as described at the beginning of Section 2 (in the Cox analyses using SAS PROC PHREG, the left truncation is handled using the counting process style of input).

In a preliminary analysis, potential nonlinearities in the effect of PM10MA on fatal MI were examined using a spline-based approach (Durrelman and Simon, 1989; Govindarajulu et al., 2007). Figure 1 shows a stepwise restricted 4-knot cubic spline graph of the relationship between PM10MA and fatal MI risk along with pointwise confidence bands. It can be seen that the dose-response relationship increases more steeply below the median of the exposure distribution, and then flattens out.

We applied the methods of this paper to test formally for a threshold effect. We took τ_{min} and τ_{max} to be equal, respectively, to the 15th and 85th percentiles of the observed distribution of PM10MA across the entire person-time experience. We first carried out analyses of the type that would be done by analyst who acted as if the threshold were known in advance. Specifically, we fit separate Cox analyses assuming τ to be known and equal, respectively, to τ_{min} , τ_{mid} , and τ_{max} . These analyses were performed using SAS PROC PHREG. We also performed a score test for each of these three values of τ . Table 4 summarizes these results. We then performed the MERT, SUP, SUP2, and SUP3 tests of $H_0 : \omega = 0$ with τ regarded as unknown. The SUP test was carried out using a 11-point grid, at equally spaced percentiles of the PM10MA distribution (ranging from the the 15th percentile to the 85th percentile). The p-value results for these tests were as follows: MERT yielded $p=0.0322$, SUP yielded $p= 0.0055$, SUP2 yielded $p=0.0164$, and SUP3 yielded $p=0.0070$. Overall, the results indicated strong evidence of a threshold effect.

6 Power Calculation

In this section, we discuss power calculation for testing a threshold. We focus here on the SUP3 statistic, although a similar development could be given for the other statistics. We restrict attention to the setting where the covariates are time-independent; in practice, most power calculations are done in this setting. In line with the classic work of Schoenfeld (1981) and Gill (1980) on power for logrank-type tests, and Hsieh and Lavori's (2000) extension of this work to power calculation for Cox regression, we consider a local inference setting in which the regression parameters β and ω associated with the covariate of interest X are small. We incorporate background covariates into the calculation; in practice, attention would typically be confined to background covariates with an known major effect, such as age at study entry.

Denote by $\tilde{\omega}$ the value of ω for which it is desired to compute the power, and suppose that the true changepoint is τ^* . Write $\tau_1 = \tau_{min}$, $\tau_2 = \tau_{mid}$, and $\tau_3 = \tau_{max}$. By arguments as in Schoenfeld (1981) and Gill (1980), we find that the joint distribution of $[W_{\tau_1}^*, W_{\tau_2}^*, W_{\tau_3}^*]$ is approximately multivariate normal with mean

$$\boldsymbol{\mu} = [\rho(\tau_1, \tau^*), \rho(\tau_2, \tau^*), \rho(\tau_3, \tau^*)]\boldsymbol{\Upsilon},$$

where $\boldsymbol{\Upsilon} = \tilde{\omega}\sqrt{nV_{\tau^*}}$, and covariance matrix \mathbf{R} given by $R_{kl} = \rho(\tau_k, \tau_l)$.

The next step is to evaluate V_{τ^*} and the various correlations. This calculation involves the covariance-like term $C(g_1, g_2)$ defined in (2.2) for various choices of g_1 and g_2 , which, in turn, involves quantities of the form $s(t, \boldsymbol{\theta}, g)$ and $s(t, \boldsymbol{\theta}, g_1, g_2)$, as defined in Section 2. In the local inference formulation, the expectations in the definitions of these quantities are taken with β and ω set equal to zero; we use the symbol E^* to denote expectations of this type. The relevant quantities are then given by $E^*[Y_i(t)h(\mathbf{W}_i, X_i)e^{\gamma^T \mathbf{W}_i}]$ for appropriate choices of the function h , and can be written as

$$E^*[Y_i(t)h(\mathbf{W}_i, X_i)e^{\gamma^T \mathbf{W}_i}] = \int f_{(X, \mathbf{W})}(x, \mathbf{w})G(t|\mathbf{w}) \exp(-\Lambda_0(t)e^{\gamma^T \mathbf{w}})e^{\gamma^T \mathbf{w}}h(\mathbf{w}, x)d\mathbf{w} dx, \quad (6.1)$$

where $f_{(X, \mathbf{W})}(x, \mathbf{w})$ is the joint density of (X, \mathbf{W}) , $G(t|\mathbf{w}) = \Pr(C \geq t|\mathbf{W} = \mathbf{w})$, and Λ_0 is the cumulative baseline hazard function. If \mathbf{W} includes discrete covariates, the integral above will actually be a mixture of integrals and summations. When the censoring does not depend on \mathbf{w} , so that $G(t|\mathbf{w}) = G(t)$, the term $G(t)$ can be taken outside the integral.

Based on (6.1), we calculate V_{τ^*} and the correlations using the formulas (2.1), (2.2), and (3.2). In practice, the calculations are done under parametric models for the joint density $f_{(X, \mathbf{W})}(x, \mathbf{w})$, the baseline hazard of the survival time, and the hazard of the censoring time; denote the respective parameter vectors by $\boldsymbol{\xi}$, $\boldsymbol{\nu}$, and $\boldsymbol{\zeta}$. In order to carry out the calculations, we need pre-study estimates of these parameters and of γ . These estimates can be obtained based on prior studies and consultation with subject-matter experts. Given these results, the critical value a for the SUP3 test is determined based on the $N(\mathbf{0}, \mathbf{R})$ distribution, and the power for a given sample size n is then obtained as $\Pr(|N(\boldsymbol{\mu}, \mathbf{R})| \geq a)$. The sample size that yields a given desired power can be found by a line search procedure, such as bisection search or the secant method.

A significant simplification ensues when the background covariates \mathbf{W} are independent of X . In this case, we have the following results.

Result A. If $g_1(x, \mathbf{w}) = \phi(x)$ and $g_2(x, \mathbf{w}) = \psi(x)$, then $C(g_1, g_2) = \mathcal{D} \text{Cov}(\phi(X), \psi(X))$, where \mathcal{D} is the probability that a given individual has an observed event, which equals

$$\mathcal{D} = E^*[\delta_i] = \iint f_{\mathbf{W}}(\mathbf{w}) G(t|\mathbf{w}) \exp(-\Lambda_0(t)e^{\boldsymbol{\gamma}^T \mathbf{w}}) e^{\boldsymbol{\gamma}^T \mathbf{w}} d\mathbf{w} dt,$$

where $f_{\mathbf{W}}(\mathbf{w})$ is the marginal density of \mathbf{W} .

Result B. If $g_1(x, \mathbf{w}) = \phi(x)$ and $g_2(x, \mathbf{w}) = \psi(\mathbf{w})$, then $C(g_1, g_2) = 0$.

Example

We illustrate these calculations in the context of the NHS study discussed in the previous section. We take X to be the baseline value of PM10MA. Time was defined as time since study entry, with age at entry handled as a sole background variable W . Baseline PM10MA was modeled using a lognormal distribution with parameters (μ, σ) ; the values used for μ and σ , the sample mean and standard deviation of log PM10MA in the NHS dataset, were $\mu = 0.90$ and $\sigma = 0.23$. The age at entry was modeled using a uniform distribution over the interval $[\mathcal{L}, \mathcal{U}]$ with $\mathcal{L} = 48$ years and $\mathcal{U} = 73$ years. We posited independence between the baseline PM10MA and the age at entry. The dependence of the survival time on W was modeled using an exponential regression model with hazard $\lambda(t|w) = \lambda_0 \exp(\gamma w)$. Censoring was modeled using a parametric proportional hazards model with regression terms involving W and W^2 and a Weibull baseline hazard: $\lambda^*(t|w) = \eta \lambda_0^* (\lambda_0^* t)^{\eta-1} \exp(\gamma_1^* w + \gamma_2^* w^2)$. In these models, age at entry was centered at 63.5 years. The model parameters, as estimated by maximum likelihood from the NHS data with time expressed in years from study entry, were $\lambda_0 = 0.001180$, $\gamma = 0.0109$, $\lambda_0^* = 0.06195$, $\eta = 6.0512$, $\gamma_1^* = 0.07585$, and $\gamma_2^* = 0.01072$. The above models were found to fit the NHS data well.

The maximum follow-up time t_{max} was set at 12 years, slightly shorter than that in the original study. We took τ_{min} to be the 5th percentile, τ_{mid} to be the 50th percentile, and τ_{max} to be the 95th percentile of the PM10MA distribution. We considered a sample size of $n=95,000$, similar to the NHS data. The $\tilde{\omega}$ values for which power calculations were performed were -0.15 , -0.65 , and -1.30 , similar to the estimates obtained in the data analysis in the preceding section.

Applying Results A and B above and the formula (3.2), we obtained the following projections: $\rho(\tau_1, \tau_2) = 0.5975$, $\rho(\tau_1, \tau_3) = 0.1659$, and $\rho(\tau_2, \tau_3) = 0.4372$. In addition, for various values of τ^* , we obtained the projections of V_{τ^*} and $\rho(\tau_j, \tau^*)$ listed in columns 2-5 of Table 5. For a test of two-sided Type I error of 0.05, a calculation based on the trivariate normal distribution leads to a critical value of $a = 2.3560$. The power results, obtained by another trivariate normal calculation, are shown in the remaining columns of Table 5. For $\tilde{\omega} = -0.15$, the power is low for all threshold values. For $\tilde{\omega} = -1.30$, is just under 90% when the threshold is at the 10th percentile and greater than 99% for all other threshold values. For $\tilde{\omega} = -0.65$, the power is greater than 90% when the threshold is at the 50th or 70th percentile, in the 70-80% range when the threshold is at the 30th or 90th percentile, and 32% when the threshold is at the 10th percentile.

7 Summary

We have discussed the problem of testing for a changepoint in the Cox survival model, which arises commonly in epidemiology. We considered both the case where the prospective changepoint is known and the case where it is not. For the latter case, we considered the MERT approach of Gastwirth (1966, 1985) and the supremum (SUP) approach of Davies (1977, 1987). We carried out a simulation study to compare the power of these two approaches in a range of settings, and we presented methodology for power calculation

at the study design stage. Matlab software for performing the tests and power calculation is available at <http://pluto.huji.ac.il/~mszucker/thresh.zip> .

Both the MERT and the SUP approaches involve a specification of a range in which the true changepoint is anticipated to lie. In general, the two approaches behave similarly. When the range specified for the changepoint is narrow, the MERT statistic offers some power advantage; when the range is wide, the SUP-type statistics are more powerful. When the range is narrow, it is possible to achieve power very similar to that of the optimal score statistic that would be used if the true changepoint were known. When the range is wide, it is not possible to get as close to the optimal power. The key quantity determining the power performance of the tests, as seen in previous papers on related problems, is the the null correlation between the score statistics for the two potential changepoint values at the extremes of the specified range. In most of the applications we have in mind, little prior information will be available on the location of the prospective changepoint. Thus, the specified range will be wide, and the correlation of the extreme score statistics moderate to low, which points in favor of the SUP approach. In general, we consider the best overall choice of test statistic to be the SUP3 statistic, which provides power that is comparable or better than the other options, and is easy to compute. In many epidemiological applications, there is reason to expect the changepoint to lie near one of the extremes, in which case the SUP2 statistic is preferable. The power calculation methodology of Section 6 will be useful in study planning.

References

- Abrahamowicz, M., Schopflocher, T., Leffondre, K., du Berger, R., and Krewski, D. (2003). Flexible modeling of exposure-response relationship between long-term average levels of particulate air pollution and mortality in the American Cancer Society Study. *Journal of Toxicology and Environmental Health - Part A*, 66, 1625–1653.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Andersen, P.K., and Gill, R. (1982). Cox’s regression model for counting processes: a large sample study. *Annals of Statistics*, 10, 1100–1120.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.
- Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64, 247–254.
- Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74, 33–43.
- Daniels, M. J., Dominici, F., Samet, J. M., and Zeger, S. L. (2000). Estimating particulate matter-mortality dose-response curves and threshold levels: an analysis of daily time-series for the 20 largest US cities. *American Journal Of Epidemiology*, 152, 397–406.
- Feder, P. I. (1975). On asymptotic distribution theory in segmented regression problems - identified case. *Annals of Statistics*, 3, 49–83.
- Fleming, T. R., and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Hoboken, N.J.: John Wiley and Sons, Inc.
- Freidlin, B., Podgor, M. J., and Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics*, 55, 883–886.
- Gastwirth, J. L. (1966). On robust procedures. *Journal of the American Statistical Association*, 61, 929–948.
- Gastwirth, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association*, 80, 380–384.
- Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.
- Govindarajulu, U. S., Malloy, E. J., Ganguli, B., Spiegelman, D., Eisen, E. A. (2009). The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *International Journal of Biostatistics*, 5, issue 1, article 2 (electronic journal - the URL of the paper is <http://www.bepress.com/ijb/vol5/iss1/2>).
- Greenland, S. (1995). Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology*, 6, 450–454.
- Hsieh, F. Y., and Lavori, P. W. (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials*, 21, 552–560.

- Levy, J. I. (2003). Issues and uncertainties in estimating the health benefits of air pollution control. *Journal of Toxicology and Environmental Health - Part A*, 66, 1865–1871.
- Kosorok, M. and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics*, 35, 957–989.
- Küchenhoff, H. and Wellisch, U. (1997). Asymptotics for generalized linear segmented regression models with unknown breakpoint. *Sonderforschungsbereich, Discussion Paper* 83.
- Paciorek, C. J., Yanosky, J. D., Puett, R. C., Laden, F., and Suh, H. H. (2009). Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Annals of Applied Statistics*, 3, 370–397.
- Pope, C. A., 3rd, Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., and Heath, C. W., Jr. (1995). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine*, 151, 669–674.
- Puett, R. C., Hart, J. E., Yanosky, J. D., Paciorek, C., Schwartz, J., Suh, H., Speizer, F. E., and Laden, F. (2009). Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses Health Study. *Environmental Health Perspectives*, 117, 1697–1701.
- Roberts, S., and Martin, M. A. (2006). The question of nonlinearity in the dose-response relation between particulate matter air pollution and mortality: Can Akaike’s Information Criterion be trusted to take the right turn? *American Journal of Epidemiology*, 164, 1242–1250.
- Samoli, E., Analitis, A., Touloumi, G., Schwartz, J., Anderson, H. R., Sunyer, J., Bisanti, L., Zmirou, D., Vonk, J.M., Pekkanen, J., Goodman, P., Paldy, A., Schindler, C., and Katsouyanni, K. (2005). Estimating the exposure-response relationships between particulate matter and mortality within the APHEA multicity project. *Environmental Health Perspectives*, 113, 88–95.
- Seber, G. A. F., and Wild, C. J. (1989). *Nonlinear Regression*. New York: Wiley.
- Durrelman, S., and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8, 551–561.
- Smith, R. L., Spitzner, D., Kim, Y., and Fuentes, M. (2000). Threshold dependence of mortality effects for fine and coarse particles in Phoenix, Arizona. *Journal of the Air and Waste Management Association*, 50, 1367–1379.
- Yanosky, J., Paciorek, C., Schwartz, J., Laden, F., Puett, R., Suh, H. (2008). Spatio-temporal modeling of chronic PM10 exposure for the Nurses’ Health Study. *Atmospheric Environment*, 42, 4047–4062.
- Zheng, G., and Chen, Z. (2005). Comparison of maximum statistics for hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrics*, 61, 254–258.
- Zucker, D. M. and Yang, S. (2006). Inference for a family of survival models encompassing the proportional hazards and proportional odds models. *Statistics in Medicine*, 25, 995–1014.

Table 1
Empirical Type I Error Levels of the Tests

Scenario	β	Grid	Test	Normal-Theory Critical Value	Type I Error Level		
					$n=500$	$n=1,000$	$n=2,000$
Flat/Sloped	0	Grid A	SUP2	2.2296	0.0580	0.0590	0.0470
			SUP3	2.3380	0.0550	0.0520	0.0440
			SUP	2.4109	0.0570	0.0470	0.0440
Flat/Sloped	0	Grid B	SUP2	2.1735	0.0590	0.0440	0.0460
			SUP3	2.2119	0.0600	0.0410	0.0460
			SUP	2.2308	0.0610	0.0410	0.0460
Modest/Sloped	log(1.2)	Grid A	SUP2	2.2296	0.0510	0.0530	0.0550
			SUP3	2.3381	0.0500	0.0470	0.0450
			SUP	2.4118	0.0540	0.0480	0.0500
Flat/Sloped	log(1.2)	Grid B	SUP2	2.1738	0.0580	0.0420	0.0460
			SUP3	2.2127	0.0540	0.0480	0.0480
			SUP	2.2314	0.0530	0.0480	0.0490

Table 2
Power Results
Flat/Sloped Log Hazard
($\beta = 0, \omega = \log(1.5)$)

Grid A

		True τ Value						
		$\Phi^{-1}(0.05)^*$	$\Phi^{-1}(0.1)$	$\Phi^{-1}(0.25)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.75)$	$\Phi^{-1}(0.9)$	$\Phi^{-1}(0.95)^*$
$n = 500$	OPTIMAL	0.109	0.141	0.237	0.333	0.265	0.161	0.136
	MERT	0.082	0.125	<i>0.216</i>	<i>0.281</i>	<i>0.222</i>	<i>0.117</i>	0.064
	SUP	<i>0.086</i>	0.122	<i>0.219</i>	0.283	<i>0.222</i>	0.116	<i>0.084</i>
	SUP2	0.098	0.142	0.223	0.247	<i>0.223</i>	0.143	0.095
	SUP3	<i>0.093</i>	<i>0.131</i>	<i>0.220</i>	<i>0.293</i>	0.226	<i>0.132</i>	<i>0.088</i>
$n = 1000$	OPTIMAL	0.155	0.221	0.447	0.577	0.451	0.245	0.163
	MERT	0.092	0.171	<i>0.350</i>	<i>0.453</i>	<i>0.366</i>	<i>0.174</i>	0.095
	SUP	<i>0.110</i>	0.164	0.357	0.479	0.372	0.162	0.101
	SUP2	0.119	0.198	<i>0.338</i>	0.379	0.342	0.191	0.119
	SUP3	<i>0.113</i>	<i>0.177</i>	<i>0.348</i>	<i>0.479</i>	<i>0.364</i>	<i>0.176</i>	0.104
$n = 2000$	OPTIMAL	0.243	0.375	0.677	0.841	0.756	0.408	0.252
	MERT	0.137	<i>0.274</i>	<i>0.565</i>	<i>0.711</i>	<i>0.604</i>	0.275	0.139
	SUP	0.144	<i>0.265</i>	0.601	0.770	0.633	<i>0.309</i>	<i>0.155</i>
	SUP2	0.163	0.291	0.531	0.599	<i>0.572</i>	0.323	0.171
	SUP3	<i>0.151</i>	<i>0.275</i>	<i>0.578</i>	<i>0.767</i>	<i>0.623</i>	<i>0.314</i>	<i>0.157</i>

Grid B

		True τ Value						
		$\Phi^{-1}(0.2)^*$	$\Phi^{-1}(0.3)$	$\Phi^{-1}(0.4)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.6)$	$\Phi^{-1}(0.7)$	$\Phi^{-1}(0.8)^*$
$n = 500$	OPTIMAL	0.209	0.266	0.317	0.333	0.323	0.286	0.231
	MERT	<i>0.160</i>	0.241	0.306	<i>0.320</i>	<i>0.309</i>	0.267	0.184
	SUP	<i>0.160</i>	0.238	<i>0.291</i>	0.321	<i>0.311</i>	<i>0.253</i>	<i>0.177</i>
	SUP2	0.166	0.250	<i>0.291</i>	<i>0.316</i>	<i>0.303</i>	<i>0.248</i>	<i>0.174</i>
	SUP3	<i>0.164</i>	0.251	<i>0.290</i>	<i>0.321</i>	0.305	<i>0.255</i>	<i>0.176</i>
$n = 1000$	OPTIMAL	0.386	0.484	0.552	0.577	0.562	0.507	0.398
	MERT	0.282	0.426	0.521	0.569	0.540	0.456	0.320
	SUP	<i>0.255</i>	<i>0.414</i>	<i>0.498</i>	<i>0.529</i>	0.483	<i>0.428</i>	<i>0.299</i>
	SUP2	<i>0.265</i>	<i>0.416</i>	<i>0.484</i> <i>0.515</i>	0.489	<i>0.425</i>	<i>0.308</i>	
	SUP3	<i>0.258</i>	<i>0.411</i>	<i>0.494</i>	<i>0.532</i>	<i>0.496</i>	<i>0.421</i>	<i>0.300</i>
$n = 2000$	OPTIMAL	0.599	0.747	0.809	0.841	0.830	0.799	0.662
	MERT	<i>0.465</i>	<i>0.682</i>	0.796	0.830	0.810	0.725	<i>0.516</i>
	SUP	<i>0.468</i>	<i>0.683</i>	<i>0.789</i>	<i>0.814</i>	<i>0.796</i>	<i>0.716</i>	<i>0.546</i>
	SUP2	0.484	0.692	<i>0.779</i>	<i>0.796</i>	<i>0.783</i>	<i>0.714</i>	0.548
	SUP3	<i>0.473</i>	<i>0.694</i>	<i>0.784</i>	<i>0.812</i>	<i>0.792</i>	<i>0.715</i>	<i>0.546</i>

* indicates τ value outside the specified range

Bold value indicates highest power in among the MERT, SUP, SUP2, and SUP3 tests

Italicized values are within 10% of the highest power

Table 3
Power Results
Modest/Sloped Log Hazard
($\beta = \log(1.2)$, $\omega = \log(1.5)$)

Grid A

		True τ Value						
		$\Phi^{-1}(0.05)^*$	$\Phi^{-1}(0.1)$	$\Phi^{-1}(0.25)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.75)$	$\Phi^{-1}(0.9)$	$\Phi^{-1}(0.95)^*$
$n = 500$	OPTIMAL	0.105	0.138	0.224	0.321	0.293	0.197	0.150
	MERT	0.075	0.108	<i>0.187</i>	<i>0.256</i>	<i>0.227</i>	<i>0.133</i>	0.086
	SUP	<i>0.079</i>	0.112	0.197	0.279	0.252	<i>0.139</i>	0.092
	SUP2	0.090	0.124	<i>0.197</i>	0.241	0.238	0.160	0.108
	SUP3	<i>0.086</i>	0.123	<i>0.208</i>	<i>0.284</i>	<i>0.253</i>	<i>0.145</i>	0.102
$n = 1000$	OPTIMAL	0.134	0.212	0.386	0.566	0.503	0.294	0.209
	MERT	0.089	<i>0.153</i>	<i>0.302</i>	<i>0.444</i>	<i>0.390</i>	<i>0.218</i>	<i>0.120</i>
	SUP	0.096	<i>0.157</i>	0.322	<i>0.469</i>	0.406	<i>0.207</i>	0.112
	SUP2	0.121	0.179	<i>0.311</i>	0.384	<i>0.379</i>	0.238	0.130
	SUP3	0.112	<i>0.168</i>	<i>0.307</i>	0.484	<i>0.397</i>	<i>0.217</i>	<i>0.123</i>
$n = 2000$	OPTIMAL	0.182	0.299	0.603	0.832	0.796	0.493	0.292
	MERT	0.128	0.218	<i>0.486</i>	<i>0.709</i>	<i>0.657</i>	0.351	0.166
	SUP	<i>0.120</i>	0.225	0.536	<i>0.719</i>	0.673	<i>0.378</i>	<i>0.186</i>
	SUP2	0.130	0.249	0.452	0.556	<i>0.615</i>	0.406	0.205
	SUP3	<i>0.125</i>	<i>0.228</i>	<i>0.504</i>	0.728	<i>0.645</i>	<i>0.388</i>	<i>0.203</i>

Grid B

		True τ Value						
		$\Phi^{-1}(0.2)^*$	$\Phi^{-1}(0.3)$	$\Phi^{-1}(0.4)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.6)$	$\Phi^{-1}(0.7)$	$\Phi^{-1}(0.8)^*$
$n = 500$	OPTIMAL	0.193	0.237	0.286	0.321	0.334	0.315	0.261
	MERT	<i>0.150</i>	<i>0.217</i>	<i>0.273</i>	<i>0.306</i>	0.324	0.296	0.218
	SUP	<i>0.153</i>	<i>0.220</i>	<i>0.278</i>	0.310	<i>0.307</i>	<i>0.283</i>	<i>0.202</i>
	SUP2	<i>0.155</i>	<i>0.222</i>	0.274	<i>0.296</i>	<i>0.311</i>	<i>0.286</i>	<i>0.216</i>
	SUP3	0.152	0.220	<i>0.277</i>	<i>0.305</i>	<i>0.313</i>	<i>0.280</i>	<i>0.210</i>
$n = 1000$	OPTIMAL	0.332	0.434	0.520	0.566	0.574	0.559	0.451
	MERT	0.246	0.392	0.491	0.544	0.545	0.494	0.351
	SUP	<i>0.221</i>	<i>0.360</i>	<i>0.456</i>	<i>0.512</i>	<i>0.521</i>	<i>0.454</i>	<i>0.328</i>
	SUP2	<i>0.233</i>	<i>0.359</i>	<i>0.437</i>	0.488	<i>0.510</i>	<i>0.461</i>	<i>0.338</i>
	SUP3	<i>0.228</i>	<i>0.359</i>	<i>0.452</i>	<i>0.511</i>	<i>0.518</i>	<i>0.455</i>	<i>0.334</i>
$n = 2000$	OPTIMAL	0.507	0.686	0.790	0.832	0.845	0.819	0.730
	MERT	<i>0.381</i>	0.619	0.765	0.823	0.820	0.767	<i>0.595</i>
	SUP	<i>0.399</i>	<i>0.606</i>	<i>0.712</i>	<i>0.777</i>	<i>0.790</i>	<i>0.744</i>	<i>0.600</i>
	SUP2	0.406	<i>0.607</i>	<i>0.706</i>	<i>0.762</i>	<i>0.769</i>	<i>0.747</i>	0.612
	SUP3	<i>0.397</i>	<i>0.606</i>	<i>0.708</i>	<i>0.778</i>	<i>0.782</i>	<i>0.744</i>	<i>0.596</i>

* indicates τ value outside the specified range
 Bold value indicates highest power in among the MERT, SUP, SUP2, and SUP3 tests
 Italicized values are within 10% of the highest power

Table 4
 Results for the NHS Study of Chronic PM₁₀ Exposure in Relation to Fatal MI
 Cox Analyses Assuming Known Threshold τ

Threshold τ	Slope Parameter β			Change in Slope Parameter ω			Score Test for $\omega = 0$ p-value
	Coefficient	Std Error	p-value	Coefficient	Std Error	p-value	
τ_{min} =15th %ile	1.42559	0.49260	0.0038	-1.32015	0.51106	0.0098	0.0085
τ_{mid} =50th %ile	0.65437	0.17735	0.0002	-0.63566	0.21739	0.0035	0.0027
τ_{max} =85th %ile	0.23873	0.09494	0.0119	-0.15704	0.17357	0.3656	0.3325

Note: The analyses adjusted for calendar year, season, and US state of residence.

Table 5
 Power Calculation Results for the NHS Study of Chronic PM₁₀ Exposure in Relation to Fatal MI
 Results are for the SUP3 Test for Threshold, Working With the Baseline PM10MA Value
 Power for $n=95,000$

τ^* (%ile)	V_{τ^*}	$\rho(\tau_1, \tau^*)$	$\rho(\tau_2, \tau^*)$	$\rho(\tau_3, \tau^*)$	$\tilde{\omega} = -0.15$	$\tilde{\omega} = -0.65$	$\tilde{\omega} = -1.30$
10th	7.9839E-5	0.9859	0.6207	0.1738	0.0616	0.3218	0.8924
30th	2.4671E-4	0.7252	0.9386	0.3109	0.0890	0.7781	0.9999
50th	3.3988E-4	0.5975	1.0000	0.4372	0.1070	0.9200	1.0000
70th	4.8884E-4	0.4174	0.9153	0.5354	0.1205	0.9631	1.0000
90th	2.7935E-4	0.2439	0.6180	0.8144	0.0842	0.7246	0.9996

Figure 1. Results of spline fit of the relative risk of fatal MI as a function of PM10MA. The solid curve is the fitted spline function, while the dashed lines are 95% pointwise confidence bands.

