Journal of Statistical Planning and Inference 139 (2009) 1633 – 1642

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A case–control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model

Mei-Ling Ting Lee[a,*], G.A. Whitmore[b], Francine Laden[d,e,f], Jaime E. Hart[d,e], Eric Garshick[c,d]

[a] *Biostatistics Division, College of Public Health, Ohio State University, Columbus, OH, USA*
[b] *Desautels Faculty of Management, McGill University, Montreal, Canada*
[c] *Pulmonary and Critical Care Medicine Section, Medical Service, VA Boston Healthcare System, USA*
[d] *Channing Laboratory, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA*
[e] *Exposure, Epidemiology and Risk Program, Department of Environmental Health, Harvard School of Public Health, Boston, MA, USA*
[f] *Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA*

ARTICLE INFO

Available online 24 May 2008

Keywords:
Biostatistics
Cardiovascular disease
Death
Disease progression
Environmetrics
Epidemiology
Exposure risk
First hitting time
Health status
Healthy worker effect
Kaplan–Meier plot
Latent process
Lung cancer
Occupational health
Stochastic process
Survival analysis
Wiener process
Work environment

ABSTRACT

A case–control study of lung cancer mortality in U.S. railroad workers in jobs with and without diesel exhaust exposure is reanalyzed using a new threshold regression methodology. The study included 1256 workers who died of lung cancer and 2385 controls who died primarily of circulatory system diseases. Diesel exhaust exposure was assessed using railroad job history from the US Railroad Retirement Board and an industrial hygiene survey. Smoking habits were available from next-of-kin and potential asbestos exposure was assessed by job history review. The new analysis reassesses lung cancer mortality and examines circulatory system disease mortality. Jobs with regular exposure to diesel exhaust had a survival pattern characterized by an initial delay in mortality, followed by a rapid deterioration of health prior to death. The pattern is seen in subjects dying of lung cancer, circulatory system diseases, and other causes. The unique pattern is illustrated using a new type of Kaplan–Meier survival plot in which the time scale represents a measure of disease progression rather than calendar time. The disease progression scale accounts for a healthy-worker effect when describing the effects of cumulative exposures on mortality.

1. Introduction

Diesel exhaust is likely to be a lung carcinogen (IARC, 1989; EPA, 2002; HEI, 1995). To investigate the relationship between diesel exhaust exposure and lung cancer we conducted a case–control study (Garshick et al., 1987) and recently updated a retrospective cohort study that included 54,973 white males (Garshick et al., 1988, 2004). In both studies we describe an increased lung cancer risk in railroad workers with exposure attributable to diesel locomotives. In Lee et al. (2004), we used a first hitting time regression model to assess lung cancer risk in the workers included in a retrospective cohort study and demonstrated that the model can provide additional insight into disease progression.

Since the original analysis of these data, interest in the health risks from diesel exhaust exposure has expanded from lung cancer to other diseases. Studies such as Pope et al. (2004) have focused attention on the role of long-term exposure to particulate matter on cardiovascular mortality. Studies conducted in professional drivers have suggested associations between long term

* Corresponding author. Tel.: +1 614 293 3918; fax: +1 614 293 3937.
E-mail address: meilinglee@cph.osu.edu (M.-L.T. Lee).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.05.023

exposures to diesel and other engine exhaust and ischemic heart disease (Tuchsen and Endahl, 1999; Hannerz and Tuchsen, 2001; Bigert et al., 2004; Finkelstein et al., 2004). In this paper, we extend the model by Lee, Garshick et al. and use the threshold regression model to reanalyze the railroad worker case–control data set to assess both lung cancer mortality and mortality from circulatory system diseases. The case–control data set also includes information regarding cigarette smoking that was not available in the retrospective cohort study. Moreover, the new model and methods are important to the fields of biostatistics and epidemiology in their own right. The threshold regression model is quite different than traditional models used in these fields, such as the proportional hazards (PH) model since insight into initial health status and disease progression attributable to multiple risk factors is provided. For an overview of threshold regression, the reader is referred to Lee and Whitmore (2006).

2. The data

The case–control data set is described fully in Garshick et al. (1987) so only a brief overview is provided here. The U.S. railroad retirement board (RRB) manages the retirement system for railroad workers. To qualify for retirement benefits, railroad workers must have at least 10 years of service. Next of kin can obtain benefits only with notification of the RRB of the worker's death. This case–control data set consists of 3641 workers that included 1256 workers who died of lung cancer (ICD 162, Eighth Revision of the International Classification of Diseases (ICD-8)) between March 1, 1981 and February 28, 1982. There were 2385 controls selected from the same RRB population so their birth dates were within 2.5 years and dates of death within 31 days of case subjects and 90% of the cases had two controls. Controls were drawn randomly from among workers who had no mention of cancer on the death certificate and who did not die by suicide, accident or an unknown cause. There were 1814 who died of circulatory disease as a cause of death and 577 who died of other causes. Six subjects had primary lung cancer noted on the death certificate, but had a circulatory disease as an underlying cause of death; hence, are included as both lung cancer and circulatory system deaths. Information on smoking habits of subjects was obtained by surveying next of kin and there were 72% with a history of smoking, 11% who never smoked, and 17% with unknown smoking histories. Yearly Interstate Commerce Commission railroad job codes were available for subjects from the RRB starting from 1959 until death or retirement.

Starting mainly after World War II, the U.S. rail industry converted from steam to diesel power and by 1959, 96% of the locomotives in service were diesel powered. We considered this year to be the effective beginning of diesel exposure in this analysis and prior to 1959, exposure was considered to have been primarily attributable to steam locomotives. A detailed job history was not available prior to 1959, and job category during the steam locomotive era was based on the last job held on or before 1959. Workers with jobs with the potential for asbestos exposure during the steam era were categorized using this job code and included locomotive shop and boiler repair workers, car repair workers, and workers responsible for the maintenance of railroad structures.

Based on an industrial hygiene survey (Woskie, 1988a, b) and a review of railroad jobs (Garshick et al., 1987) diesel exhaust exposure was considered in three categories: (1) train operations personnel, such as engineers, fireman, conductors, brakemen, and hostlers, (2) railroad locomotive shop workers, such as machinists, electricians and supervisors and (3) all other employees such as clerks, ticket agents, station agents, railroad car repair workers, and maintenance of way workers. The train operations personnel would experience regular exposure attributable to operating trains. Although the greatest diesel exhaust exposures were measured in the locomotive repair shops (Woskie, 1988a, b), the job codes included in the shop category also included non-diesel locomotive shop workers and the overall degree of exposure was uncertain. Finally, the other employees would have infrequent or no exposure. For the analysis that follows, we equate diesel-exhaust exposure with employment in the trains operations category starting in 1959, subsequently referred to as the engineer–brakeman job category.

The main explanatory variables for our analysis include age at death (variable age), year the subject first joined the railroad (variable first), year of retirement from the railroad (variable retirement), years of work as an engineer or brakeman (variable exposed), years of unexposed work, i.e., not as an engineer or brakeman (variable unexposed), years in retirement before death (variable retired) and indicator variables for 'ever worked as engineer or brakeman' (variable engineer), 'ever worked in a railroad shop' (variable shopworker), 'ever smoked cigarettes' (variable smoking) and 'ever exposed to asbestos' (variable asbestos). Code 0 for the smoking indicator variable includes workers who never smoked and those with an unknown smoking status. The work history variables allow the experience of each subject to be divided into years before joining the railroad, years of work in each job category, and years of retirement before death. We measure each subject's history from birth. The years from birth until the subject joined the railroad is denoted by variable prior_railroad. Years of railroad employment from year of hire to 1959 were considered to be unexposed work. Further breakdowns of the work history are employed to refine the analysis.

3. Disease progression and threshold regression models

A subject's health status with respect to a particular disease (say, lung cancer) is a latent (unobservable) stochastic process that fluctuates until the first time when the subject's condition has deteriorated to a threshold that is the point of death. Lee and Whitmore (2004) refer to such models as first-hitting time (FHT) models. Fig. 1 shows the basic situation. The figure illustrates health status with respect to lung cancer for case and control subjects starting at birth. The zero level is the threshold. Observe that the health status of a case subject (a subject who dies of lung cancer in this instance) deteriorates until it reaches the zero threshold in 1981–1982 and, hence, the case subject dies from that cause. On the other hand, the health status of a control subject terminates above the zero level in 1981–1982 because another cause of death has intervened. Therefore, the survival time of a

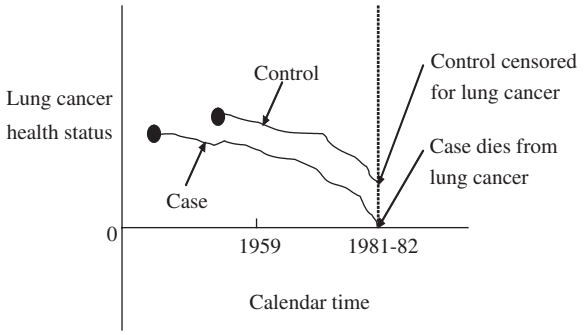

Fig. 1. Illustrative lung cancer health status paths for a case subject (lung cancer death) and a control subject (another cause of death) who died in 1981–1982. The control subject has a censored survival time with respect to lung cancer.

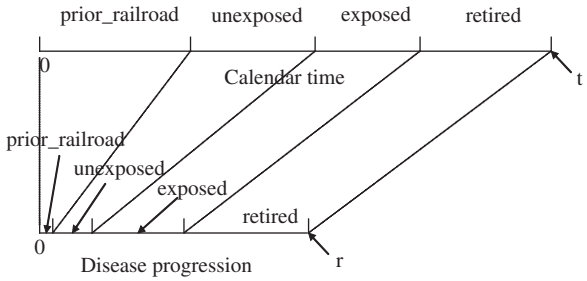

Fig. 2. Correspondence of calendar time (t) and disease progression (r) scales for a representative subject having exposed and unexposed employment intervals prior to retirement (and death). Disease progression is measured in retirement-equivalent years.

control subject is a censored observation with respect to lung cancer death. It will be seen later that, irrespective of how case and control were defined when the data set was assembled, the role of a case subject may be considered as from any specific cause, whether it is lung cancer, a cardiovascular disease or another cause.

Lee et al. (2004) introduce the concept of *disease progression*. Different calendar time intervals during life will have different rates of disease progression, depending on the disease risk and health stress to which the subject is exposed during the interval. Disease progression pertains to the advance of the particular disease under study (say, lung cancer or cardiovascular disease). The more slowly a disease progresses, the longer the subject postpones potential death from that disease. Fig. 1 shows how the health status of a subject fluctuates stochastically over time with a trend that is sometimes shallow and sometimes steep. The varying trend suggests that a time-scale transformation might be helpful. The disease progression scale is marked off in equal units of *expected health deterioration*. Thus, health status will have a constant expected rate of change or linear trend when measured against the disease progression scale. Fig. 2 illustrates the mapping of calendar time intervals into disease progression intervals for the work history of a representative subject. In this figure and later, we represent calendar time and disease progression by variables t and r, respectively. Observe in this example how the rate of disease progression varies for different intervals of life experience. The interval corresponding to time before joining the railroad, represented by variable prior_railroad, involves a slow rate of disease progression, presumably because the worker is young, healthy and living in a benign environment. In contrast, the worker's employment interval in an exposed environment, experienced after joining the railroad and represented by variable exposed, involves more rapid disease progression per unit of calendar time. Likewise, disease progresses at different rates during intervals of unexposed employment and retirement. Arbitrarily, we choose retirement as the reference environment and measure disease progression in units of *retirement-equivalent years*. Thus, the interval for retirement, represented by variable retired, maps into an interval of disease progression of the same length because we mark off the disease progression scale in retirement-equivalent years.

Following the line of reasoning in Lee et al. (2004), the link between elapsed calendar time t_i and disease progression r_i for a particular subject i has the following relationship in this study.

$$r_i = \sum_{j=1}^{J} \alpha_j c_j^{(i)}(t_i). \tag{1}$$

The mathematical relationship imitates the visual representation in Fig. 2. The general setup assumes J job categories. Experience prior to joining the railroad and retirement are treated as two job categories, specifically categories 1 and J. Notation $c_j^{(i)}(t_i)$ represents the time spent by subject i in job category j in calendar interval $[0, t_i]$. Thus, $\sum_j c_j^{(i)}(t_i) = t_i$. Disease progression per

unit time in job category j is an unknown parameter α_j that will be estimated. The vector of parameters $\alpha_j, j = 1, \ldots, J-1$, in (1) is denoted subsequently by $\boldsymbol{\alpha}$. The Jth job category (retirement) is chosen as the reference category. Thus, the rate α_J for the Jth job category is set to unity (i.e., $\alpha_J = 1$). This specification implies that the unit of measurement on the disease progression scale corresponds to one retirement-equivalent year.

Observe from both Fig. 2 and formula (1) that the same disease progression value r_i will be given to all work histories that yield the same employment intervals $c_j^{(i)}(t_i)$, irrespective of whether these employment intervals arise early or late in the work history or arise in fragments. It is the total time spent in each kind of job that matters.

4. Sample log-likelihood function

As in Lee et al. (2004), we take the latent health status process, defined on the disease progression scale, to be a Wiener diffusion process (i.e., Brownian motion with linear drift). The first-hitting-time for such a process follows an inverse Gaussian distribution. The cited article gives a number of arguments for the appropriateness of using a Wiener process to describe health status. The data analyzed here also lend support to the reasonableness of this theoretical model, as is shown later.

The inverse Gaussian distribution for the first-hitting-time depends on the mean parameter μ of the underlying Wiener process and on the initial or starting health status level $X(0) = x_0$. The variance parameter of the underlying Wiener process is set to unity because the health status process is latent and, hence, can be given an arbitrary measurement unit. The parameter μ measures the rate of decline in health status for each retirement-equivalent year of disease progression and is assumed to be a constant rate. The constant value of μ implies that the health status of a subject will tend to have a linear trajectory with reference to the disease progression scale. The second parameter x_0 measures the subject's health at birth and, hence, is the distance that health status must fall for the subject to die. In engineering parlance, x_0 may be viewed as a subject's initial physiological strength and μ as his expected rate of decline in strength measured on the retirement-equivalent time scale.

We let $f(r|\mu, x_0)$ and $F(r|\mu, x_0)$ denote, respectively, the probability density function (p.d.f.) and cumulative distribution function (c.d.f.) of the first-hitting-time distribution defined in terms of disease progression r. Both parameters μ and x_0 will be linked to k regression covariates that will be represented by row vector $\boldsymbol{z} = (1, z_1, \ldots, z_k)$. The leading 1 in $\boldsymbol{z}$ allows for a constant term in the regression relationship. An identity function of form

$$\mu = \boldsymbol{z}\boldsymbol{\beta} = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k$$

is used to link parameter μ to the covariates and a logarithmic function

$$\ln(x_0) = \boldsymbol{z}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_k z_k$$

is used to link parameter x_0 to the covariates. Here $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_k)'$, where β_0 and γ_0 are regression constants. The disease progression parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J-1})'$ will enter the sample log-likelihood function in the logarithmic form $\ln(\alpha_j)$ to ensure positive parameter estimates.

Lee et al. (2004) derived likelihoods for threshold models for cohort studies. In this article, we consider likelihoods for threshold models for case–control studies. Let the numbers of case and control subjects be denoted by n_1 and n_2, respectively. We use $\mu^{(i)}$ and $x_0^{(i)}$ to denote μ and x_0, respectively, for subject i. Each case subject contributes an observed lifetime from the reference date to the year of death. Each control subject contributes a censored survival time (censored by another cause of death) measured from the same reference date. Hence, each case subject i contributes probability density $f(r_i|\mu^{(i)}, x_0^{(i)})$ to the sample likelihood function, for $i = 1, \ldots, n_1$, and each control subject i contributes survival probability $\overline{F}(r_i|\mu^{(i)}, x_0^{(i)}) = 1 - F(r_i|\mu^{(i)}, x_0^{(i)})$ to the sample likelihood function, for $i = n_1 + 1, \ldots, n_1 + n_2$.

Observe that the functions are defined in terms of disease progression r. The sample log-likelihood function to be maximized therefore has the form:

$$\ln L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n_1} \{\ln f(r_i|\mu^{(i)}, x_0^{(i)})\} + \sum_{i=n_1+1}^{n_1+n_2} \{\ln \overline{F}(r_i|\mu^{(i)}, x_0^{(i)})\}. \tag{2}$$

The form of the likelihood function requires some justification. We recall that a cohort study and case–control study differ in the way subjects are chosen (see, for example, Rothman, 1986). In the former, subjects are chosen based on varying exposure to a risk factor of interest and are tracked over time until an endpoint of interest occurs (e.g., death). In a case–control study, subjects are chosen on the basis of the presence (case) or absence (control) of one particular endpoint of interest and their exposures to a risk factor are compared. In both study designs, the association of exposure to endpoint is of interest. The concept of latent survival times is convenient in the context of competing causes of death that is involved in this application. In a cohort study, a subject is imagined to have latent survival times $\{S_1, \ldots, S_C\}$ for C causes of death, measured from time of birth or some other reference age. The observed cause of death d and observed survival time S_d are given by

$$S_d = \min\{S_c,\ c = 1, \ldots, C\}. \tag{3}$$

If follow-up is limited, then the survival time S_d may be censored. Furthermore, the latent survival time for any cause of death other than the observed cause is also censored because $S_c > S_d$ for all $c \neq d$. In contrast, in our case–control study, all subjects who

die of some cause in 1981–1982 are imagined to have had latent survival times $\{S_1, \ldots, S_C\}$ for C causes of death. The observed cause of death d and observed survival time S_d are again given by (3). The outcome S_d is the observed time lapse from the reference date until the year of death (1981–1982). Again, the latent survival time for any cause of death other than the observed cause is a censored observation because $S_c > S_d$ for all $c \neq d$. As a Wiener process is a time-reversible stochastic process, the formulation of the case–control model, which looks backward from the year of death, is equivalent to the forward-looking representation for a cohort study. Both give rise to the same sample log-likelihood expression given in (2). The foregoing formulation remains unchanged when a disease progression scale replaces the calendar time scale.

As well as justifying the mathematical form of the sample likelihood function, we wish to elaborate on the limitations of the inferences we are drawing in this case–control sampling context. As the description of the RRB data set in Section 2 makes clear, the sample contains all lung cancer deaths during 1981–1982 and one or two randomly matched controls from the same source of deaths. Subjects whose death certificates mentioned death from cancer, suicide, accident or an unknown cause were excluded as controls. Control subjects were matched to case subjects on birth date and time of death. Thus, the *reference population* for statistical inferences here is a population of workers whose health and lifestyle characteristics and occupational exposures correspond to those of U.S. railroad workers in the RRB group who died in 1981–1982, with the exception of the excluded causes of death. The setup of the sample likelihood function in (2) takes no account of the population proportions of case and control subjects in the RRB data set and we do not attempt to estimate these proportions. The subpopulation of workers with lung cancer deaths (the cases) is accounted for in the study. Thus, the part of the sample likelihood function dealing with cases yields inferences that are valid for lung cancer deaths in the reference population. The controls are a sample from a *matching subpopulation* of workers. These workers are matched to the cases in the manner just described and therefore are representative of workers who die from included causes of death, other than lung cancer, with birth and death dates that match the lung cancer subpopulation. Thus, the part of the sample likelihood function dealing with controls yields valid inferences for workers dying of included causes in the matching subpopulation. This subpopulation is precisely the population of interest in a case–control study where it is desired to assess differential outcomes for subjects who have matching characteristics except for those of exposure.

5. Results

The threshold regression model was fitted to the case–control data by maximizing the sample log-likelihood function in (2). Indicator covariates smoking, engineer and asbestos were used as explanatory variables for log-initial health status $\ln(x_0)$ and mean rate parameter μ. The effect of adding the indicator variable shopworker will be postponed until later. It was expected (and later confirmed) that the length of exposure has some effect. For this reason, the variable exposed was broken into intervals shorter and longer than 10 years, creating variables exposed-under 10 and exposed-over 10. The latter variable measures the excess of exposure beyond 10 years. Keep in mind that the intervals in question do not necessarily represent continuous employment in the exposed job category, although for almost all subjects the employment interval is continuous. The short and long intervals of employment without exposure did not show similar large differences so variable unexposed was not partitioned. The threshold regression model was fitted first with lung cancer deaths treated as cases, next with cardiovascular deaths treated as cases and finally deaths from other causes (neither lung cancer nor cardiovascular disease) as cases. Tables 1–3 show the results, which we now interpret.

1. *Healthy-worker effect*: The phrase *healthy worker effect* describes the phenomenon of employed persons appearing to be healthier than unemployed persons who are comparable on other characteristics. Thus, the simple fact of being employed

Table 1
Lung cancer deaths form the cases

Parameter	Variable	Estimate	*P*-value
$\ln(x_0)$			
	Engineer	1.15681	0.000
	Smoking	0.08389	0.000
	Asbestos	0.08202	0.000
	Constant	2.40497	0.000
μ			
	Engineer	−0.89114	0.000
	Smoking	−0.16693	0.000
	Asbestos	−0.07824	0.000
	Constant	−0.27750	0.000
$\ln(\alpha_j)$			
$j = 1$	prior_railroad	−2.78791	0.000
$j = 2$	Exposed-under 10 yr	−0.33369	0.001
$j = 3$	Exposed-over 10 yr	0.08839	0.014
$j = 4$	Unexposed years	−1.37027	0.000

Covariates smoking, engineer and asbestos are retained for $\ln(x_0)$ and μ. Exposure intervals are divided between short ('under 10' years) and long ('over 10' years).

Table 2
Cardiovascular deaths form the cases

Parameter	Variable	Estimate	*P*-value
$\ln(x_0)$			
	Engineer	0.67089	0.000
	Smoking	−0.08175	0.001
	Asbestos	0.14787	0.000
	Constant	2.78009	0.000
μ			
	Engineer	−0.56060	0.000
	Smoking	0.09355	0.000
	Asbestos	−0.12960	0.000
	Constant	−0.64107	0.000
$\ln(\alpha_j)$			
$j=1$	prior_railroad	−4.04036	0.000
$j=2$	Exposed-under 10 yr	−0.86899	0.000
$j=3$	Exposed-over 10 yr	0.11498	0.001
$j=4$	Unexposed years	−0.98792	0.000

Covariates engineer, smoking and asbestos are retained for $\ln(x_0)$ and μ. Exposure intervals are divided between short ('under 10' years) and long ('over 10' years).

Table 3
Deaths from causes other than lung cancer and cardiovascular disease form the cases

Parameter	Variable	Estimate	*P*-value
$\ln(x_0)$			
	Engineer	0.91553	0.000
	Smoking	0.11826	0.000
	Asbestos	0.04706	0.101
	Constant	2.61406	0.000
μ			
	Engineer	−0.70169	0.000
	Smoking	−0.10049	0.001
	Asbestos	−0.05413	0.067
	Constant	−0.33419	0.000
$\ln(\alpha_j)$			
$j=1$	prior_railroad	−2.98270	0.000
$j=2$	Exposed-under 10 yr	−0.17049	0.167
$j=3$	Exposed-over 10 yr	0.02736	0.624
$j=4$	Unexposed years	−1.19148	0.000

Covariates smoking, engineer and asbestos are retained for $\ln(x_0)$ and μ. Exposure intervals are divided between short ('under 10' years) and long ('over 10' years).

is associated with better health. Adjustments for this effect are essential in studies of occupational risk. We now look at the evidence for these effects in the study.

The estimated coefficients $\ln(\alpha_j)$ show that years prior to joining the railroad, represented by variable prior_railroad, are clearly healthy, for subjects dying of lung cancer, cardiovascular disease or other causes. For lung cancer cases, for example, Table 1 gives a log-coefficient of −2.78791 for prior_railroad. This coefficient corresponds to $\hat{\alpha}_1 = \exp(-2.78791) = 0.062$, which implies that disease progression prior to joining the railroad runs at a youthful rate that is only 6% of the rate during retirement, which acts as the reference (i.e., $\alpha_J = 1$). The employment interval unexposed also shows a strong healthy-worker effect relative to retirement for all three causes of death. The healthy-worker effect for exposed-under 10 varies with cause of death. For lung cancer and other causes of death, the log-coefficient for this variable is negative but not large in absolute value. In the case of other causes of death, the coefficient is not significantly different from zero. Thus, for these two causes of death, the healthy worker effect appears to be weak for short employment as an engineer or brakeman. For cardiovascular deaths (Table 2), the two coefficients for exposed-under 10 years and unexposed are almost identical, both being large and negative. For example, for cardiovascular deaths, the rate coefficient for unexposed is $\hat{\alpha}_4 = \exp(-0.98792) = 0.372$, which means that the disease progression rate during unexposed employment is only 37% of that during retirement. Finally, for exposures in excess of 10 years, as captured by variable exposed-over 10 years, the log-coefficients for deaths from lung cancer, cardiovascular disease and other causes are all positive, although only significantly so for the first two causes. These coefficients show a complete absence of a healthy worker effect and imply, in fact, disease progression at rates at or exceeding those in retirement.

2. *Effects of exposure*: Now we turn to consider the regression coefficients of indicator variables smoking, engineer and asbestos for the parameters $\ln(x_0)$ and μ. All of the indicator effects are significantly different from zero, except for those of asbestos for other causes of death (*P*-values of 0.101 and 0.067, respectively). Those for engineer are largest in absolute magnitude for

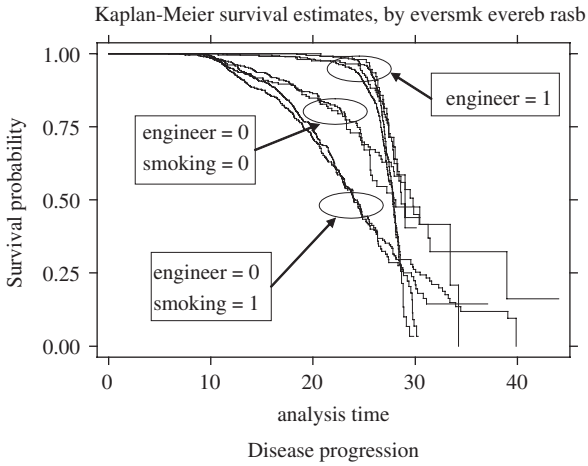

Fig. 3. Kaplan–Meier plot of survival probability against disease progression for lung cancer deaths.

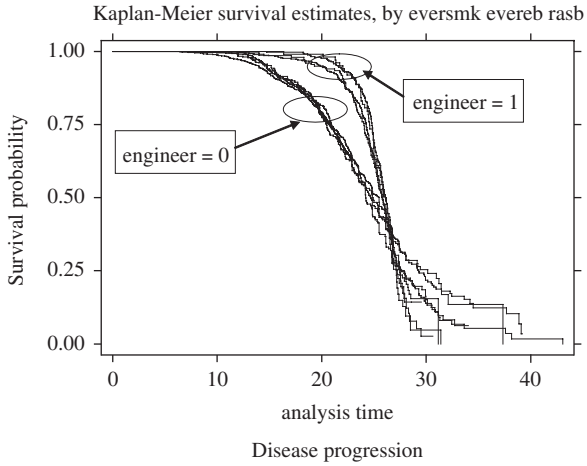

Fig. 4. Kaplan–Meier plot of survival probability against disease progression for cardiovascular deaths.

all causes of death. In fact, the engineer coefficients are almost an order of magnitude larger than those for the other indicator variables.

The fact that the coefficients for each indicator variable have opposite signs for $\ln(x_0)$ and μ in all cases needs explanation. For example, the effects for engineer in the case of lung cancer deaths (Table 1) are 1.15681 for $\ln(x_0)$ and -0.89114 for μ. These coefficients imply that workers who have been engineers or brakemen have a higher initial health status x_0 but a steeper rate of decline (i.e., more negative value of μ) than workers who have never been engineers or brakemen. The expected survival time $E(S)$ for the threshold regression model is given by $E(S) = x_0/|\mu|$, measured in retirement-equivalent years. Thus, the effects of engineer on the two parameters x_0 and μ offset each other in the numerator and denominator of the formula, as both parameters are larger when the worker has been an engineer or brakeman. The net result of these offsetting effects on estimated mean survival times will be examined in the Discussion section.

3. *Insights from modified KM plots*: In seeking a cleaner summary of the findings, we have prepared three sets of Kaplan–Meier (KM) survival plots, one set for each cause of death. Each plot has eight curves corresponding to all possible outcomes of the indicator variables smoking, engineer and asbestos. Because our model has introduced the concept of disease progression as a time scale, we plot estimated survival probability against disease progression, which in this case is denominated in retirement-equivalent years. Thus, the time scales of the survival plots are adjusted for the subject's work history in terms of healthy-worker effects and other variability in the progress of disease. To our knowledge, this kind of plot is new and, in this application, shows striking contrasts between subjects who were engineers or brakemen (engineer = 1) and those who were not (engineer = 0), for all three causes of death. Figs. 3–5 show the plots. The plots are semi-parametric because although

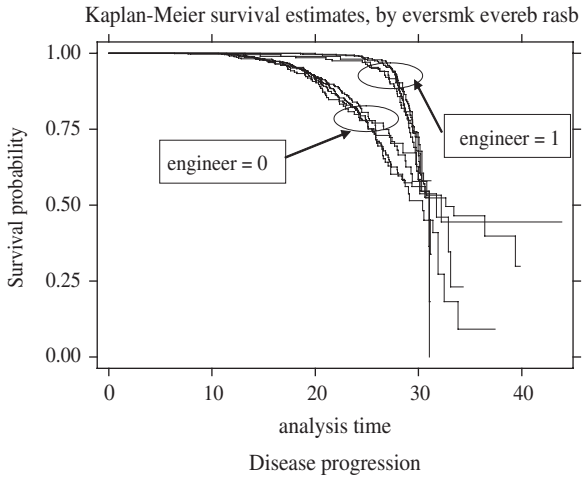

Fig. 5. Kaplan–Meier plot of survival probability against disease progression for other deaths.

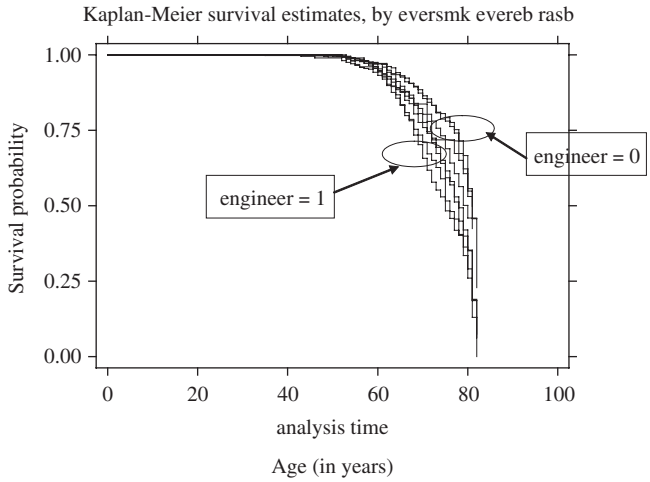

Fig. 6. Kaplan–Meier plot of survival probability against age (in years) for lung cancer deaths.

the disease progression scale has parameters estimated from the threshold regression model, the KM plots themselves are nonparametric estimates of survival distributions.

The KM plots show distinct clusters of survival curves for subjects who were engineers or brakemen (engineer = 1) for all three disease categories. The survival curves for subjects dying of lung cancer who were not in the engineer–brakemen job category show secondary clusters based on whether they smoked or not. The survival curves for subjects dying of cardiovascular disease or other causes who were not in the engineer–brakeman job category show less separation based on smoking but pull apart in the right tails (a feature that is not highlighted in the figures). For subjects in the engineer–brakemen job category (and presumably highly exposed to diesel exhaust), the KM plots show high survival probabilities over a long initial range of disease progression and then a sharp decline toward zero as disease progression reaches the neighborhood of 30 retirement-equivalent years. This feature is found in the KM plots for engineer = 1 for all three causes of death.

To contrast the modified KM plots with conventional KM plots in which survival probability is plotted against age, we have prepared Fig. 6. This plot shows the conventional KM plot of survival probability against age (in years) for lung cancer deaths. The plot has eight curves corresponding to all possible outcomes of the indicator variables engineer, smoking, and asbestos. Observe that the survival curves in Fig. 6 are roughly ordered in two groups by the variable engineer with the engineer–brakeman job category engineer = 1 having poorer survival prospects (more leftward survival curves). This conventional KM plot does not reveal the distinctive shape and separation of the survival curves seen in the corresponding modified KM plot for lung cancer shown in Fig. 3.

6. Discussion

1. *Effects of shop employment*: A total of 665 workers had worked in a railroad shop at some point of time (evershop = 1). Of these, 17 had also worked in the engineer–brakeman job category at some time. The indicator variable shopworker was added to the regression models displayed in Tables 1–3. The regression coefficients of this indicator variable for parameters $\ln(x_0)$ and μ are tabulated below. Its effects on $\ln(x_0)$ are negative for all three causes but significant only for lung cancer and cardiovascular disease. Its effects on μ are all positive but are significant at the conventional 0.05 level only for lung cancer.

Parameter	$\ln(x_0)$		μ	
	Estimate $\hat{\gamma}_j$	P-value	Estimate $\hat{\beta}_j$	P-value
Lung cancer	−0.19269	0.000	0.07744	0.028
Cardiovascular disease	−0.12689	0.002	0.05568	0.132
Other	−0.10315	0.111	0.04102	0.423

   Introduction of covariate shopworker to the three regression models leaves the other regression coefficients essentially unchanged (the results are not shown). Thus, on balance, shop workers appear to experience some increased risk of death from both lung cancer and cardiovascular disease.
2. *Mean survival time*: Tables 1–3 showed that the effects of the indicator variables engineer, smoking, and asbestos have opposite signs for parameters $\ln(x_0)$ and μ. The effects for variable engineer, for example, are largest and imply higher values for initial health status x_0 and more negative values for μ. As noted earlier, the expected survival time $E(S)$ for subjects (measured in retirement-equivalent years) is related to the model parameters as follows: $E(S) = x_0/|\mu|$. Thus, the effects of engineer on the two parameters are offsetting in the numerator and denominator as both parameters rise with exposure.
   Panels (a), (b) and (c) of Table 4 show the estimated mean survival times, calculated from the fitted regression models, for all $2^3 = 8$ possible outcomes of the three indicator variables smoking, engineer and asbestos for the three diseases. The means are denominated in retirement-equivalent years. Panel (d) shows the number of subjects for each of the eight outcome combinations. The two rows in each disease panel compare the estimated means for engineer = 0 and engineer = 1, respectively, for all combinations of the other two indicator variables. Although the log-coefficients for engineer and the modified KM plots show distinct survival patterns, depending on whether engineer = 1 or engineer = 0, the estimated mean survival times show mixed comparisons. For example, for lung cancer deaths, non-smoking subjects who were never exposed to asbestos (smoking = 0 and asbestos = 0) have an estimated mean survival time $E(S)$ of 39.9 and 30.1 retirement-equivalent years when engineer = 0 and engineer = 1, respectively. A clear decrement in life expectancy with exposure is evident. The corresponding numbers for cardiovascular deaths are 25.1 and 26.2 retirement-equivalent years, respectively. Note that the latter two estimates for cardiovascular deaths are close, suggesting that the offsetting effects come close to cancelling each other with respect to mean survival time (in retirement-equivalent years) in this instance.
3. *Pattern of the KM plots*: The study raises interesting questions about the causes of the peculiar pattern of survival seen in the modified KM plots. Certainly, within the first decade of exposure (i.e., during the interval exposed-under 10), subjects in the engineer–brakemen job category enjoy a good measure of the healthy-worker effect. This effect is completely erased, however, beyond 10 years (as measured by covariate exposed-over 10) when they experience a high rate of disease progression. Moreover, as parameter μ is large (and negative) for engineers and brakemen, health status declines rapidly with every retirement-equivalent year of living. Yet, because these workers start from an elevated initial level of health status x_0, the rapid decline does not cause death until after an initial delay. Thus, in spite of their high rate of decline in health status, workers who were most exposed to diesel exhaust have a stay of death because they are inherently healthier to start. Nevertheless, when death finally comes, it happens to almost all of them in rapid succession (at about 30 retirement-equivalent years of survival). These observations suggest an exposure model where there is a latent period of exposure that may initiate disease. Any of the diseases, once initiated, generally proceed very quickly and predictably. The suggestion in the data that workers attracted to the engineer–brakeman job category are initially healthier than other workers is an interesting conjecture that needs further study.
4. *Comparison with proportional hazards models*: We noted in the introduction that the threshold regression model is quite different than traditional models used in these fields, such as the proportional hazards (PH) model. It is evident from the plots in Figs. 3–5 that the proportional hazards assumption is not appropriate for the groups of survival curves corresponding to engineer = 1 and engineer = 0 because the curves can be seen to cross over each other. This crossing property is inconsistent with proportional hazards. Nonetheless, if PH regression were applied in this context, it would tend to overlook the effects that our threshold regression model has detected. On the other hand, a comparison of inverse Gaussian c.d.f.s (implied by the Wiener diffusion threshold regression model) with the modified KM plots show good agreement and, hence, support the use of the model. We do not present these comparative plots here.
5. *Validity*: Lee et al. (2004), in their cohort study of railroad workers, described the elevated risk of lung-cancer death for subjects employed in the engineer/brakeman job category. A similar association was observed using the case–control data set.

Table 4
(a)–(c) Estimated mean survival times in retirement-equivalent years, classified by cause of death, and (d) numbers of subjects, for all combinations of indicator variables engineer, smoking and asbestos

Engineer	Smoking = 0		Smoking = 1	
	Asbestos = 0	Asbestos = 1	Asbestos = 0	Asbestos = 1
(a) Lung cancer				
Engineer = 0	39.9	33.8	27.1	25.0
Engineer = 1	30.1	30.7	28.7	29.4
(b) Cardiovascular disease				
Engineer = 0	25.1	24.3	27.1	25.4
Engineer = 1	26.2	27.5	26.2	27.2
(c) Other cause				
Engineer = 0	40.9	36.9	35.4	33.0
Engineer = 1	32.9	32.8	33.8	33.8
(d) Numbers of subjects				
Engineer = 0	467	291	1062	630
Engineer = 1	205	104	543	339

This re-analysis also supports the findings of the original case–control analysis, and has extended the findings in ways that give greater insight into the effects of diesel exposure, especially with respect to risks from cardiovascular diseases.

6. *Collapsible lifetime regression model*: The lifetime model presented here and in Lee et al. (2004) is a special case of a class of lifetime regression models referred to in the literature as *collapsible models*. These models have evolved from theoretical and practical suggestions put forward by several authors including, most recently, Oakes (1995), Kordonsky and Gertsbakh (1997), Duchesne and Lawless (2000), and Duchesne and Rosenthal (2003). The practical contexts for much of this development work have been various degradation and internal wear processes for equipment. The analogy to degradation of human health is immediate, where the engineering role of cumulative equipment usage corresponds to our disease progression measure here.

Acknowledgements

This research is supported in part by NIH Grants OH008649 (Lee), CA79725 and CCR115818 (Garshick), and the Natural Sciences and Engineering Research Council of Canada (Whitmore).

References

Bigert, C., Klerdal, K., Hammar, N., Hallqvist, J., Gustavsson, P., 2004. Time trends in the incidence of myocardial infarction among professional drivers in Stockholm 1977–1996. Occup. Environ. Med. 61, 987–991.
Duchesne, T., Lawless, J., 2000. Alternative time scales and failure time models. Lifetime Data Anal. 6, 157–179.
Duchesne, T., Rosenthal, J.S., 2003. On the collapsibility of lifetime regression models. Adv. Appl. Probab. 35, 755–772.
Finkelstein, M.M., Verma, D.K., Sahai, D., Stefov, E., 2004. Ischemic heart disease mortality among heavy equipment operators. Am. J. Ind. Med. 46 (1), 16–22.
Garshick, E., Schenker, M.B., Munoz, A., Segal, M., Smith, T.J., Woskie, S.R., Hammond, S.K., Speizer, F.E., 1987. A case–control study of lung cancer and diesel exhaust exposure in railroad workers. Amer. Rev. Respir. Dis. 135, 1242–1248.
Garshick, E., Schenker, M.B., Munoz, A., Segal, M., Smith, T.J., Woskie, S.R., Hammond, S.K., Speizer, F.E., 1988. A retrospective cohort study of lung cancer and diesel exhaust exposure in railroad workers. Amer. Rev. Respir. Dis. 137 (4), 820–825.
Garshick, E., Laden, F., Hart, J.E., Rosner, B., Smith, T.J., Dockery, D.W., Speizer, F.E., 2004. Lung cancer in railroad workers exposed to diesel exhaust. Environ. Health Perspect. 112, 1539–1543.
Hannerz, H., Tuchsen, F., 2001. Hospital admissions among male drivers in Denmark. Occup. Environ. Med. 58 (4), 253–260.
Health Effects Institute, 1995. Diesel exhaust. A critical analysis of emissions, exposure, and health effects. Diesel Working Group, Health Effects Institute.
IARC (International Agency for Research on Cancer), 1989. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 46. IARC, Lyons.
Kordonsky, K.B., Gertsbakh, I., 1997. Multiple time scales and the lifetime coefficient of variation: engineering applications. Lifetime Data Anal. 3, 139–156.
Lee, M.-L.T., Whitmore, G.A., 2004. First hitting time models for lifetime data. In: Rao, C.R., Balakrishnan, N. (Eds.), Handbook of Statistics: Advances in Survival Analysis, vol. 23. pp. 537–543.
Lee, M.-L.T., Whitmore, G.A., 2006. Threshold regression for survival analysis: modeling event times by a stochastic process. Statist. Sci. 21, 501–513.
Lee, M.-L.T., Garshick, E., Whitmore, G.A., Laden, F., Hart, J., 2004. Assessing lung cancer risk to railroad workers using a first hitting time regression model. Environmetrics 15, 1–12.
Oakes, D., 1995. Multiple time scales in survival analysis. Lifetime Data Anal. 1, 7–18.
Pope, C.A., Burnett, R.T., Thurston, G.D., Thun, M.J., Calle, E.E., Krewski, D., Godleski, J.J., 2004. Cardiovascular mortality and long-term exposure to particulate air pollution. Circulation 109, 71–77.
Rothman, K.J., 1986. Modern Epidemiology. Little Brown and Company, Boston, MA.
Tuchsen, F., Endahl, L.A., 1999. Increasing inequality in ischaemic heart disease morbidity among employed men in Denmark 1981–1993: the need for a new preventive policy. Int. J. Epidemiol. 28 (4), 640–644.
US Environmental Protection Agency, 2002. Health assessment document for diesel engine exhaust. Washington, DC.
Woskie, S.R., et al., 1988a. Estimation of the diesel exhaust exposures of railroad workers: I. Current exposures. Amer. J. Ind. Med. 13 (3), 381–394.
Woskie, S.R., et al., 1988b. Estimation of the diesel exhaust exposures of railroad workers: II. National and historical exposures. Amer. J. Ind. Med. 13 (3), 395–404.