CAM

Centre for Applied Microeconometrics

Institute of Economics
University of Copenhagen

http://www.econ.ku.dk/CAM/

Bounds in Competing Risks Models and the War on Cancer

Bo E. Honoré
&
Adriana Lleras-Muney

2004-08

The activities of CAM are financed by a grant from
The Danish National Research Foundation

Bounds in Competing Risks Models and the War on Cancer*

Bo E. Honoré† Adriana Lleras-Muney‡

This Version: April 2004

Abstract

Competing risks models are fundamentally unidentified. This paper derives bounds for aspects of the underlying distributions under a number of different assumptions. These bounds are then applied to mortality data from the US. We find that trends in cancer show much larger improvements than was previously estimated.

This paper is still very incomplete and preliminary. Please do not circulate or quote without permission. Future versions will be available at www.princeton.edu/~honore.

Keywords: Bounds, Competing Risks, Cancer.

JEL Classification:

1 Introduction

In 1971 President Nixon declared war on cancer. As a result the Nixon administration created a National Cancer Program administered by the National Cancer Institute, and increased the federal funds allocated to cancer research dramatically.[1] Thirty years later, however, many have declared this war a failure (Bailar and Smith (1986), Bailar and Gornik (1997), etc). Overall cancer statistics

*We would like to thank the seminar participants at CAM at the University of Copenhagen, MIT, and the Harvard–MIT–Boston University Health seminar for their suggestions. This research was supported by the National Institute on Aging, Grant Number K12-AG00983 to the National Bureau of Economic Research (Adriana Lleras–Muney) and by National Science Foundation, The Gregory C. Chow Econometric Research Program at Princeton University, and the Danish National Research Foundation, through CAM at The University of Copenhagen (Bo Honoré).

†Mailing Address: Department of Economics, Princeton University, Princeton, NJ 08544-1021. Email: honore@Princeton.edu.

‡Mailing Address: Woodrow Wilson School of Public and International Affairs, 320 Wallace Hall, Princeton University, Princeton, New Jersey 08544-1013. Email: alleras@princeton.edu

[1]The National Cancer Institute's budget is approximately $4.3 billion (or 18% of the budget for the NIH).

confirm this view: age-adjusted incidence rates and mortality rates show a bleak picture. From 1970 to 1994, age-adjusted mortality from cancer increased by 6 percent (Bailar and Gornik (1997)). Incidence rates also increased during the same period (Howe, Wingo, Thun, Ries, Rosenberg, Feigal, and Edwards (2001)).

At the same time, age-adjusted mortality rates from cardiovascular disease have fallen quite dramatically. (See Figure 1a.) It has been hypothesized that the decline in mortality rates from cardiovascular disease is somewhat responsible for the rise in cancer mortality. In other words, perhaps if there had been no progress in cardiovascular disease, we might have observed different trends in cancer mortality. The intuition behind the hypothesis that observed cancer trends are biased is that the fall in mortality rates from cardiovascular disease leaves more and perhaps different individuals at risk for cancer. Indeed for younger individuals, for whom cardiovascular disease is not a large competing risk, there have been large improvements in cancer: since 1973, cancer mortality for children and adolescents (under age 20) has fallen by more than 50% across all types of cancers, and it fell by 20% for young adults ages 20 to 44 (Doll (1991)). Interestingly these reductions have occurred in spite of the increases in cancer incidence for both groups (see figure 1b). The same is not true for older adults. Although it has long been recognized that dependent competing risks can affect trends in cancer mortality, no estimates of cancer trends exist that account for this possibility. In fact in 1990, the Extramural Committee to Assess Measures of Progress Against Cancer recommended "additional research on how cancer statistics are affected by changes in other causes of death."

This paper derives bounds for aspects of the underlying distributions under a number of different assumptions. Most importantly, we do not assume that cancer and cardiovascular disease are independent risks, and impose very weak parametric assumptions in order to obtain identification. The theoretical contribution of the paper is to provide a framework to estimate competing risk models with interval data and discrete explanatory variables, both of which are common in empirical applications. This framework is then applied to mortality data from the US to estimate the trends in cancer mortality, which are the most widely used measure of overall progress against cancer.[2]

[2]There are several measures used to assess progress in cancer, including age-adjusted incidence rates, 5 year survival rates conditional on diagnosis, and mortality rates. Both survival rate conditional on diagnosis and incidence rates are affected by improvement in diagnosis technology. Better diagnostic tools allow for detection of tumors at earlier stages, generating a mechanical increase in survival rates that does not reflect improvements in prevention or treatment. Similarly improved detection increases observed incidence, even though disease rates may not have changed. For example, incidence rates for prostate cancer have more than doubled since 1974, and there has been a 50% increase in the 5-year survival from prostate cancer. However there has been no change in mortality from prostate cancer at any age. Studies suggest that the improvements in survival are mostly attributable to earlier

We find that trends in cancer show much larger improvements than previously estimated.

Formally, a competing risks model is a duration model where the observed duration is the shortest of a number of latent durations, as well as its identity

$$(T, \delta) = (\min \{T_1, T_2, ..., T_K\}, \arg\min \{T_1, T_2, ..., T_K\}).$$

See, for example, Kalbfleisch and Prentice (1980) or Crowder (2001). Much of the terminology in this literature is motivated by medical applications where T_k could be the unobserved duration until death from a specific cause such as cancer or cardiovascular disease, T the observed duration until death and δ the cause of death.

There are a number of economic applications of the competing risks model in economics. For example, Flinn and Heckman (1982) investigated the duration of unemployment where an employed individual could terminate a spell of unemployment either by finding a job or by leaving the labor market. Katz and Meyer (1990) used the competing risks model to study the probability of leaving unemployment through recalls and new jobs. Other applications include studies of age at marriage or cohabitation (Berrington and Diamond (2000)), Ph. D. completion (Booth and Satchell (1995)), and mortgage termination (Deng, Quigley, and Van Order (2000)). The competing risks model is also closely related to the Roy (1951) model studied in Heckman and Honoré (1990) and Heckman, Smith, and Clements (1997).

2 Competing Risks

In order to simplify the exposition, we will focus on the case where $K = 2$ in what follows. The general case requires no additional ideas, but the notation is substantially more cumbersome in that case.

The identification of the competing risks model is tricky. The key result in this literature is that for any joint distribution of (T_1, T_2), there exists (unique) univariate distribution for S_1 and S_2, such that if S_1 and S_2 are independent, then the distribution of $(\min \{T_1, T_2\}, \delta)$ equals that of $(\min \{S_1, S_2\}, \arg\min \{S_1, S_2\})$ (see Cox (1962) and Tsiatis (1975)). Since this exercise can be carried out conditional on a set of explanatory variables X, the relationship between (T_1, T_2) and X is fundamentally unidentified. On the other hand, the conditional distribution for S_1 and S_2

detection made possible by the introduction of PSA screening in the late 1980s (Welch, Schwartz, and Woloshin (2000)). Additionally, diagnosis is a function of access to care, further complicating the interpretation of changes in incidence and 5-year survival rates. For these reasons, when reporting to the Senate Appropriations committee in 1990 the Extramural Committee to Assess Measures of Progress against Cancer concluded that age-specific cancer mortality is the best measure of progress again cancer.

(conditional on X) that is constructed by imposing conditional independence of S_1 and S_2, is often "unreasonable" in the sense that it violates the assumptions that one would be willing to impose a priori. For example, when studying mortality by cause, one may be willing to assume that a drug X affects only S_1 but by imposing independence we will estimate that drug X also affects S_2. More generally, imposing independence, when the risks are indeed dependent, will yield biased estimates of the cause-specific hazard rates and of the effect of covariates on those hazards (Slud and Byar (1988)). Additionally, in a heterogenous population, competing risks will generally not be independent even if the risks are independent for every individual in that population (Vaupel and Yashin (1999)). The following example illustrates this point.

Example 1 *Suppose that there are two causes of death labelled 1 and 2 and that there are two types of individuals. For each type, w, the individual's durations until death are independent with hazard for the j'th cause of death*

$$\lambda_j(t\,|w) = \theta_j^w$$

If $P(w=1) = p$ and $P(w=2) = (1-p)$ and if one assumes that the unobserved durations until death are independent (conditional on w) then the implied hazards would be

$$\lambda_j(t) = \frac{p\theta_j^1 e^{-\left(\theta_1^1+\theta_2^1\right)t} + (1-p)\,\theta_j^2 e^{-\left(\theta_1^2+\theta_2^2\right)t}}{pe^{-\left(\theta_1^1+\theta_2^1\right)t} + (1-p)\,e^{-\left(\theta_1^2+\theta_2^2\right)t}}$$

Now imagine that originally,

$$\begin{aligned}\theta_1^1 &= \theta_2^1 = 1\\ \theta_1^2 &= \theta_2^2 = 2\end{aligned}$$

and that in a later period

$$\begin{aligned}\theta_1^1 &= 1 \qquad \theta_2^1 = \frac{1}{2}\\ \theta_1^2 &= 2 \qquad \theta_2^2 = 1\end{aligned}$$

In words, the hazard for the second cause of death is reduced by a factor of 2. Furthermore assume that $p = \frac{1}{2}$. In this example, the implied hazard for the first cause of death, assuming independence, would be

$$\lambda_1(t) = \frac{e^{-2t} + 2e^{-4t}}{e^{-2t} + e^{-4t}}$$

in the original period, and

$$\lambda_j(t) = \frac{e^{-3t/2} + 2e^{-3t}}{e^{-3t/2} + e^{-3t}}$$

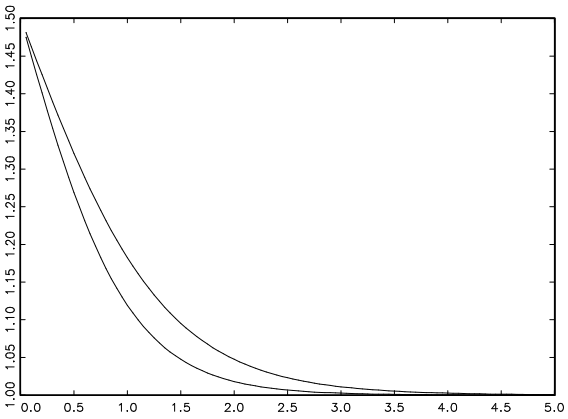


Figure 1: Hazard Rates for the Example

in the later period. The two hazard rates are depicted in figure 1. In this example, there was no change in the distribution for the duration until the first cause of death, but the dependence between the two risks makes it seem as if its hazard has increased.

There are a number of examples like this in the literature, and they suggests that it might be fruitful to ask what features of the conditional distribution of (T_1, T_2), given some explanatory variable X, can be identified if one is willing to impose restrictions on those conditional distributions. For example, in the example above, one might ask what one could say about the change in the distribution of T_2 if one is willing to assume that there was no change in the marginal distribution of T_1.

Heckman and Honoré (1989) show (essentially) that with a mixed proportional hazard model or an accelerated failure time model on the marginal distributions of T_1 and T_2, the full model is identified if one is willing to assume that the support of the effect of X on the hazard functions for T_1 and T_2 is $\Re^2_+$. See for example Crowder (2001) for a discussion of restrictiveness of this assumption. A recent paper by Abbring and van den Berg (2003) relaxes these conditions somewhat by showing that the unbounded support assumption can be dispensed with if one is willing to make additional assumptions. However, in many situations the covariates of interest have bounded support and are not continuous. For example, analyses of mortality use data from death certificates, which contain demographic information that is all categorical, such as race, gender and marital status. Moreover, the proofs in Heckman and Honoré (1989) and Abbring and van den Berg (2003) rely crucially on the duration, T, being observed exactly, whereas in many data sets durations are observed in groups. This raises the question of what can be learned in competing risks models if one is willing to impose restrictions that are weaker than those in Heckman and Honoré (1989) and Abbring and van den Berg (2003). It turns out that many realistic assumptions will result in

models for which the parameters of interest are not point–identified. The discussion in the rest of the paper will therefore focus around the identified region for a parameter.

Competing risks models are a subset of sample selection models. The research presented here is therefore closely related to the literature on bounds in sample selection models (see for example Manski (1990)), although the results here take advantage of the special structure of the competing risks model.

3 Bounds in Some Specific Competing Risks Models

We have

$$T^* = \min\{T_1, T_2\}, \qquad \delta = 1\{T_1 < T_2\}$$

where we are interested in features of the distribution of (T_1, T_2) given X.

Following, for example Prentice and Gloeckler (1978) and Meyer (1990), we assume that (T_1, T_2) has a continuous positive density conditional on X, but that T^* is grouped so we observe events like (T, δ, X), where $T = t_k$ if $t_k < T^* \leq t_{k+1}$ for $k = 1, .., M$ and $t_{M+1} = \infty$. In the following we assume M is finite, so that there is only a finite number of possible outcomes. We also assume that δ is unobserved when $T^* > t_M$. In other words, we allow T^* to be censored at t_M.

3.1 Marginal distributions. No assumptions.

Following the approach of, for example, Manski (2003), it is straightforward to generate bounds on the marginal distributions of T_1 and T_2. Since[3] $T < T_1 \leq T^+ \cdot \delta + \infty \cdot (1 - \delta)$, where T^+ is the upper endpoint of the interval in which T belongs, one can bound the α^{th} quantile of the distribution of T_1 (given X) by

$$\left\{F \in \mathcal{F} : Q_\alpha(T|X) \leq F^{-1}(\alpha) \leq Q_\alpha\left(T^+ \cdot (1 - \delta) + \infty \cdot \delta \middle| X\right)\right\}$$

where $\mathcal{F}$ is the set of all distributions functions that satisfy whatever restrictions one might impose. When $\mathcal{F}$ is the set of all distribution functions (and when there is no grouping), these bounds are given in Peterson (1976), who also provides bounds on the joint distribution of T_1 and T_2.

3.2 The effect of explanatory variables with parametric restrictions.

The nonparametric bounds above can be quite wide (see the numerical example in Peterson (1976)). The main methodological contribution of the research presented here is to show how parametric

[3]Here, $0 \cdot \infty$ is defined to equal 0.

assumptions can help tighten the bounds on the object of interest in unidentified competing risks models. In each of the examples, we will use the fact that for any distribution of (T_1, T_2) given X, there exist an observationally equivalent discrete distribution for which the probability of a tie is 0. This follows from the fact that only a discretized version of T is observed. If X can take a finite number of values, this means that for all the cases we consider, there will be an observationally equivalent case in which the vector of all the random variables has a discrete distribution with a finite number of points of support.

The leading example considered in the paper is one in which the explanatory variable, X, has a multiplicative effect on T_1,

$$T_1 = \begin{cases} S_1 & \text{for } X = 0 \\ aS_1 & \text{for } X = 1 \end{cases}$$

where the multiplicative effect, a, is the main object of interest. This model is an example of an accelerated failure time in which the support conditions in Heckman and Honoré (1989) and Abbring and van den Berg (2003) are not satisfied.

We will consider two versions of this model: (a) X also has a multiplicative effect on T_2,[4] and (b) no assumption is made on the effect of X on T_2. In each case, we make use of the fact that for any parameter value which is consistent with the observed distribution of the data, there is a discrete distribution of the underlying random variables that makes it consistent with the data. In asking whether a particular value of a is consistent with the observed distribution of the data, there is therefore no loss in generality by assuming that the underlying distributions are discrete (with support that depends on a). The points of support will be denoted by s, and the associated probabilities by p.

3.3 Case (a): X has multiplicative effect on T_2.

We consider first the case where X has a multiplicative effect on both of the latent distributions,

$$(T^*, I) = \begin{cases} (\min\{S_1, S_2\}, 1\{S_1 < S_2\}) & \text{for } X = 0 \\ (\min\{\alpha S_1, \beta S_2\}, 1\{\alpha S_1 < \beta S_2\}) & \text{for } X = 1 \end{cases}, \tag{1}$$

where (S_1, S_2) is independent of X.

[4] This is a version of the location–shift model (in logarithmic scale) considered by Lin, Robins, and Wei (1996), except that they assume that T_1 is observed whether or not it exceeds T_2.

In this case, the relevant probabilities are

$$P\left(t < S_1 < t+1, S_1 < S_2\right) \tag{2}$$

$$P\left(t < S_2 < t+1, S_2 < S_1\right) \tag{3}$$

$$P\left(t < \alpha S_1 < t+1, \alpha S_1 < \beta S_2\right) = P\left(\frac{t}{\alpha} < S_1 < \frac{t+1}{\alpha}, S_1 < \frac{\beta}{\alpha} S_2\right) \tag{4}$$

$$P\left(t < \beta S_2 < t+1, \beta S_2 < \alpha S_1\right) = P\left(\frac{t}{\beta} < S_2 < \frac{t+1}{\beta}, S_2 < \frac{\alpha}{\beta} S_1\right) \tag{5}$$

Now consider the set of numbers $\{0, 1, 2, 3, ..., t_{Max}\} \cup \left\{0, \alpha^{-1}, 2\alpha^{-1}, 3\alpha^{-1}, ..., t_{Max}\alpha^{-1}\right\}$. Label this set $\{q_1, q_2, ..., q_K\}$. These are the relevant numbers as far as the marginal distribution of T_1 is concerned. Also consider the set of numbers $\{0, 1, 2, 3, ..., t_{Max}\} \cup \left\{0, \beta^{-1}, 2\beta^{-1}, 3\beta^{-1}, ..., t_{Max}\beta^{-1}\right\}$. Label this set $\{r_1, r_2, ..., r_L\}$. These are the relevant numbers for the marginal distribution of T_2.

The first two graphs in Figure 2 depict the events in (2) and (3), and in (4) and (5), respectively. The dashed lines in the graphs corresponds to the numbers $\{0, 1, 2, 3, ..., t_{Max}\}$ and the dotted lines to $\left\{0, \alpha^{-1}, 2\alpha^{-1}, 3\alpha^{-1}, ..., t_{Max}\alpha^{-1}\right\}$ and $\left\{0, \beta^{-1}, 2\beta^{-1}, 3\beta^{-1}, ..., t_{Max}\beta^{-1}\right\}$.

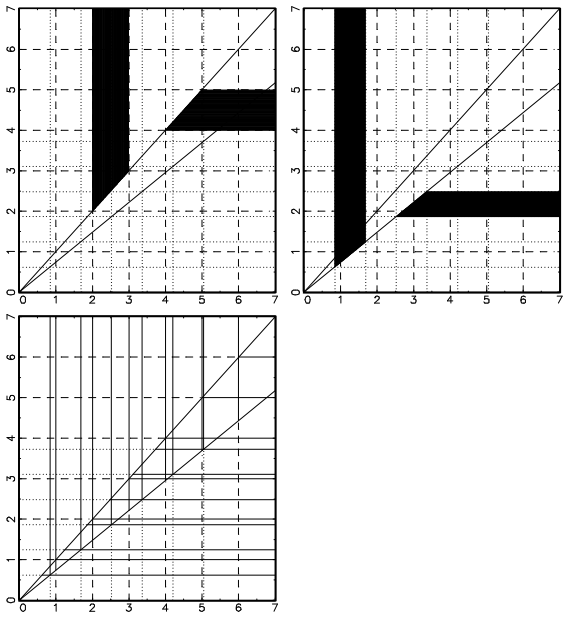


Figure 2: Illustration of Points of Support for Case (a)

It is clear that the probabilities of those events would be unchanged if one redistributed probability within each of the polygons depicted (in solid lines) in the third graph. There is therefore no loss of generality in assuming that the distribution of (S_1, S_2) is discrete, with one point of support in each of the regions.

The identified region for (α, β) is the set of (a, b) such that there exists $p(s_1, s_2)$ satisfying

$$\sum_{\substack{t_k < s_1 < t_k + 1 \\ s_2 > s_1}} p(s_1, s_2) = P(T = t_k, I = 1 | X = 0), \tag{6}$$

$$\sum_{\substack{t_k < s_2 < t_k + 1 \\ s_1 > s_2}} p(s_1, s_2) = P(T = t_k, I = 0 | X = 0), \tag{7}$$

$$\sum_{\substack{t_k < as_1 < t_k + 1 \\ bs_2 > as_1}} p(s_1, s_2) = P(T = t_k, I = 1 | X = 1), \tag{8}$$

$$\sum_{\substack{t_k < bs_2 < t_k + 1 \\ as_1 > bs_2}} p(s_1, s_2) = P(T = t_k, I = 0 | X = 1) \tag{9}$$

$$\sum_{s_1, s_2} p(s_1, s_2) = 1, \tag{10}$$

$$p(s_1, s_2) \geq 0 \tag{11}$$

(where the first four equations hold for all $k = 1, .., M$).

These equations have exactly the same structure as the constraints of a linear programming problem. Analogous to Honoré and Tamer (2003), one can check whether a feasible solution to such a linear programming problem exists for a given a and b by solving an auxiliary linear programming problem and checking whether its optimal value is 0 (the alternative being that it is negative). We will show that as in Honoré and Tamer (2003), one can consistently estimate the identified region for (α, β), by maximizing the optimal value in the sample analogs to the auxiliary linear programming problem.

Specifically, for given a and b consider the linear programming problem

$$f(a, b) = \max \sum -v_i \tag{12}$$

subject to

$$\begin{aligned}
v_k + \sum_{\substack{t_k<s_1<t_k+1 \\ s_2>s_1}} p(s_1,s_2) &= P(T=t_k, I=1|\, X=0) \qquad k=1,\dots M, \\
v_{M+k} + \sum_{\substack{t_k<s_2<t_k+1 \\ s_1>s_2}} p(s_1,s_2) &= P(T=t_k, I=0|\, X=0\ ) \qquad k=1,\dots M, \\
v_{2M+k} + \sum_{\substack{t_k<as_1<t_k+1 \\ bs_2>as_1}} p(s_1,s_2) &= P(T=t_k, I=1|\, X=1\ ) \qquad k=1,\dots M, \\
v_{3M+k} + \sum_{\substack{t_k<bs_2<t_k+1 \\ as_1>bs_2}} p(s_1,s_2) &= P(T=t_k, I=0|\, X=1\ ) \qquad k=1,\dots M, \\
v_{4M+1} + \sum_{s_1,s_2} p(s_1,s_2) &= 1, \\
p(s_1,s_2) &\geq 0 \qquad \text{for all } (s_1,s_2), \\
v_i &\geq 0 \qquad k=1,\dots 4M+1
\end{aligned}$$

This linear programming problem has a feasible solution:

$$\begin{aligned}
v_k &= P(T=t_k, I=1|\, X=0) \qquad k=1,\dots M, \\
v_{M+k} &= P(T=t_k, I=0|\, X=0\ ) \qquad k=1,\dots M, \\
v_{2M+k} &= P(T=t_k, I=1|\, X=1\ ) \qquad k=1,\dots M, \\
v_{3M+k} &= P(T=t_k, I=0|\, X=1\ ) \qquad k=1,\dots M, \\
v_{4M+1} &= 1, \\
p(s_1,s_2) &= 0 \qquad \text{for all } (s_1,s_2)
\end{aligned}$$

and the optimal function value in (12) is 0 if the equations (6)–(11) have a solution and it is strictly negative otherwise.

By mimicking the argument in Honoré and Tamer (2003), it is easy to establish that $\widehat{f}(a,b)$ converges uniformly to $f(a,b)$ where the former has been defined by the same linear programming problem but with all the probabilities, P, replaced by consistent estimates. However, the situation here is different from that considered in Honoré and Tamer (2003) as the objective function here is piecewise constant.

Lemma 1 *$f(a,b)$ is piecewise constant over a finite number of regions.*

With this, it follows that

Theorem 2 *Define the function* $\widehat{f}$ *by the linear programming problem above, but with the probabilities in the constraints,* P*, replaced by consistent estimators. The set of maximizers of* $\widehat{f}(a,b)$ *is set–consistent for the identified region for* (α,β)*.*

Note that unlike for example Manski and Tamer (2002) and Honoré and Tamer (2003), we do not need to define the estimator to be the set of parameter values, (a,b), such that $\widehat{f}(a,b) \geq \max \widehat{f} - \varepsilon_n$ where ε_n is some sequence to be chosen. This is due to the discontinuity of the objective function established in Lemma 1.

Imposing $b = 1$ in this example, will give the identified region for a, under the exclusion restriction that X has no effect on T_2.

Remark 1. The competing risks model considered in this section is a special case of the kind of general sample selection models that have been considered in the econometric literature. Specifically, if the durations are not grouped, then one can write the model in (1) as a switching regression model. See Amemiya (1985). Specifically, consider $\log(T_1)$

$$\log(T_1) = X \cdot \log(a) + \varepsilon_1$$

where $\log(T_1)$ *is observed only if*

$$X \cdot (\log(b) - \log(a)) + (\varepsilon_2 - \varepsilon_1) < 0$$

The standard sufficient conditions for identification of such models require that X *has "full rank" conditional on the probability that the selection criterion is satisfied (i.e., conditional on the so–called propensity score). See for example Ahn and Powell (1993). This sufficient condition is not satisfied here. Moreover, it is clear that a model with a finite number of points of support and a discrete outcome will not be point identified (by the same intuition as to why a semiparametric discrete choice model is not identified if the explanatory variables take only a finite number of values).*

3.4 Case (b): no assumption is made on the effect of X on T_2.

It is also relatively straightforward to establish bounds for a in the case where one makes no assumption on the effect of X on T_2. Specifically, suppose that

$$(T^*, I) = \begin{cases} (\min\{S_1, S_2\}, 1\{S_1 < S_2\}) & \text{for } X = 0 \\ \left(\min\left\{\alpha S_1, \widetilde{S}_2\right\}, 1\left\{\alpha S_1 < \widetilde{S}_2\right\}\right) & \text{for } X = 1 \end{cases},$$

where $\left(S_1, S_2, \widetilde{S}_2\right)$ is independent of X. The identified region for α is the set of a's such that there exist $p(s_1, s_2)$ and $\widetilde{p}(s_1, s_2)$ satisfying

$$\begin{aligned}
\sum_{\substack{t_k < s_1 < t_k + 1 \\ s_2 > s_1}} p(s_1, s_2) &= P(T = t_k, I = 1 | X = 0), \\
\sum_{\substack{t_k < s_2 < t_k + 1 \\ s_1 > s_2}} p(s_1, s_2) &= P(T = t, I = 0 | X = 0) \\
\sum_{\substack{t_k < a s_1 < t_k + 1 \\ s_2 > a s_1}} \widetilde{p}(s_1, s_2) &= P(T = t_k, I = 1 | X = 1), \\
\sum_{\substack{t_k < s_2 < t_k + 1 \\ a s_1 > s_2}} \widetilde{p}(s_1, s_2) &= P(T = t, I = 0 | X = 1) \\
\sum_{s_1, s_2} p(s_1, s_2) &= 1, \\
\sum_{s_1, s_2} \widetilde{p}(s_1, s_2) &= 1, \\
\sum_{s_2} p(s_1, s_2) &= \sum_{s_2} \widetilde{p}(s_1, s_2) \\
p(s_1, s_2) &\geq 0, \qquad \widetilde{p}(s_1, s_2) \geq 0
\end{aligned}$$

where the last set of equality constraints captures the constraint that the marginal distribution of S_1 should be the same whether it is calculated from the distribution of (S_1, S_2) or from the distribution of $\left(S_1, \widetilde{S}_2\right)$. These equations again have the structure of the constraints of a linear programming problem.

As in section 3.3, one can estimate the identified region as a set of maximizers of a function that is defined as the optimal function value for a linear programming problem.

3.5 Counterfactuals

The explanatory variable, X, is often a time–dummy. In that case, it natural to ask what the distribution of T would have been like if only the distribution of T_1 had changed.

Consider for example the setup on section 3.3 and define

$$\widetilde{T}^* = \min\{\alpha S_1, S_2\}$$

This is the duration that one would observe if X has the hypothesized effect on the first latent duration but has no effect on the second duration. This could then be compared to the distribution of T^* given $X = 0$ in order to find the effect that X has on T through T_1 alone.

Unfortunately, such an exercise is not literally possible if T^* is grouped. In that case one can only get the distribution of the grouped version of T^* given $X = 0$. It is therefore natural to also consider the distribution of the grouped version of $\widetilde{T}^*$. This is the equivalent of considering the distribution function for $\widetilde{T}^*$ at $t_1, t_2, ..$

For a given α and β and a given for $p(\cdot, \cdot)$ we have

$$\begin{aligned} P\left(\widetilde{T}^* < t_k\right) &= P\left(\min\left\{\alpha S_1, S_2\right\} < t_k\right) \\ &= P\left(\alpha S_1 < t_k, S_2 < t_k\right) \\ &= P\left(S_1 < t_k/\alpha, S_2 < t_k\right) \\ &= \sum_{s_1 < t_k/\alpha, s_2 < t_k} p\left(s_1, s_2\right) \end{aligned}$$

It is important to note that this is not affected by the fact that the points of support are not uniquely determined and the non–uniqueness of the location of the points in the polygons in the third graph of figure 2 does not change whether $s_1 < t_k/\alpha$, $s_2 < t_k$.

One can therefore calculate population bounds on $P\left(\widetilde{T}^* < t_k\right)$ by minimizing and maximizing (over a and b) the function $\sum_{s_1 < t_k/\alpha, s_2 < t_k} p\left(s_1, s_2\right)$ subject to (6)–(11). Unfortunately, the sample analog of this will not produce a consistent estimator of the upper and lower bounds on $P\left(\widetilde{T}^* < t_k\right)$. The reason is that there is no guarantee that the sample version of (6)–(11) will have a solution for any value of a or b.

It is also not possible to estimate the upper and lower bounds by referring to the solution to (12). The reason for this is that for given (a, b), the solution for $p(\cdot, \cdot)$ need not be unique. However, this suggests constructing consistent estimators for the upper and lower bounds as follows. Let $\widehat{\Theta}$ be the set of maximizers of

$$f(a, b) = \max \sum -v_i \tag{13}$$

subject to

$$\begin{aligned}
v_k + \sum_{\substack{t_k<s_1<t_k+1 \\ s_2>s_1}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=1|\, X=0) \qquad k=1,\dots M,\\
v_{M+k} + \sum_{\substack{t_k<s_2<t_k+1 \\ s_1>s_2}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=0|\, X=0\,) \qquad k=1,\dots M,\\
v_{2M+k} + \sum_{\substack{t_k<as_1<t_k+1 \\ bs_2>as_1}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=1|\, X=1\,) \qquad k=1,\dots M,\\
v_{3M+k} + \sum_{\substack{t_k<bs_2<t_k+1 \\ as_1>bs_2}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=0|\, X=1\,) \qquad k=1,\dots M,\\
v_{4M+1} + \sum_{s_1,s_2} p(s_1,s_2) &= 1,\\
p(s_1,s_2) &\geq 0 \qquad \text{for all } (s_1,s_2),\\
v_i &\geq 0 \qquad k=1,\dots 4M+1
\end{aligned}$$

and let $\widehat{f}$ be the optimal function value. The consistent estimators of the upper bound on $P\left(\widetilde{T}^* < t_k\right)$ is then obtained by maximizing $g(a,b)$ over (a,b) in $\widehat{\Theta}$ where

$$g(a,b) = \max \sum_{s_1<t_k/a, s_2<t_k} p(s_1,s_2)$$

subject to

$$\begin{aligned}
v_k + \sum_{\substack{t_k<s_1<t_k+1 \\ s_2>s_1}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=1|\, X=0) \qquad k=1,\dots M,\\
v_{M+k} + \sum_{\substack{t_k<s_2<t_k+1 \\ s_1>s_2}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=0|\, X=0\,) \qquad k=1,\dots M,\\
v_{2M+k} + \sum_{\substack{t_k<as_1<t_k+1 \\ bs_2>as_1}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=1|\, X=1\,) \qquad k=1,\dots M,\\
v_{3M+k} + \sum_{\substack{t_k<bs_2<t_k+1 \\ as_1>bs_2}} p(s_1,s_2) &= \widehat{P}(T=t_k, I=0|\, X=1\,) \qquad k=1,\dots M,\\
v_{4M+1} + \sum_{s_1,s_2} p(s_1,s_2) &= 1,\\
\sum -v_i &= \widehat{f}\\
p(s_1,s_2) &\geq 0 \qquad \text{for all } (s_1,s_2),\\
v_i &\geq 0 \qquad k=1,\dots 4M+1
\end{aligned}$$

The consistent estimators of the lower bound on $P\left(\widetilde{T}^* < t_k\right)$ is obtained by minimizing $g(a,b)$ over (a,b) in $\widehat{\Theta}$ where

$$g(a,b) = \min \sum_{s_1 < t_k/a, s_2 < t_k} p(s_1, s_2)$$

subject to the same constraints.

3.6 Exclusion Restrictions

Exclusion restrictions are sometimes useful in improving identification. One way to model an exclusion restriction in the competing risks model is to assume that the explanatory variable X is independent of one of the latent durations

$$(T^*, I) = \begin{cases} (\min\{S_1, S_2\}, 1\{S_1 < S_2\}) & \text{for } X = 0 \\ \left(\min\left\{\widetilde{S}_1, S_2\right\}, 1\left\{\widetilde{S}_1 < S_2\right\}\right) & \text{for } X = 1 \end{cases},$$

This model generalizes the competing risks model considered by, for example, Faraggi and Korn (1996), and it is in the spirit of many econometric models in which exclusion restrictions are used to obtain point–identification.

In this section, we will discuss how to obtain bounds on difference in the distribution functions for S_1 and $\widetilde{S}_1$. This is essentially done as in section 3.1, but with the added restriction that the marginal distribution for S_2 is the same in the two subsamples given by $X = 0$ and $X = 1$.

Suppose that we are interested in bounding $P(T_1 \leq t \,|X = 1) - P(T_1 \leq t \,|X = 0)$ for some t. In this case, the relevant points of support are given in figure 3.[5]

The lower bound for $P(T_1 \leq t \,|X = 1) - P(T_1 \leq t \,|X = 0)$ is then the value of

$$\min \sum_{s_1 \leq t} \widetilde{p}(s_1, s_2) - \sum_{s_1 \leq t} p(s_1, s_2)$$

subject to

$$\sum_{\substack{t_k < s_1 < t_k + 1 \\ s_2 > s_1}} p(s_1, s_2) = P(T = t_k, I = 1 |\, X = 0\,), \qquad \sum_{\substack{t_k < s_2 < t_k + 1 \\ s_1 > s_2}} p(s_1, s_2) = P(T = t, I = 0 |\, X = 0\,)$$

$$\sum_{\substack{t_k < s_1 < t_k + 1 \\ s_2 > s_1}} \widetilde{p}(s_1, s_2) = P(T = t_k, I = 1 |\, X = 1\,), \qquad \sum_{\substack{t_k < s_2 < t_k + 1 \\ s_1 > s_2}} \widetilde{p}(s_1, s_2) = P(T = t, I = 0 |\, X = 1\,)$$

$$\sum_{s_1, s_2} p(s_1, s_2) = 1, \qquad \sum_{s_1, s_2} \widetilde{p}(s_1, s_2) = 1, \qquad \sum_{s_1} p(s_1, s_2) = \sum_{s_1} \widetilde{p}(s_1, s_2)$$

$$p(s_1, s_2) \geq 0, \qquad \widetilde{p}(s_1, s_2) \geq 0$$

[5] Figure 3 is drawn for the case where the observations are censored after 9 periods and $t = 5.5$.

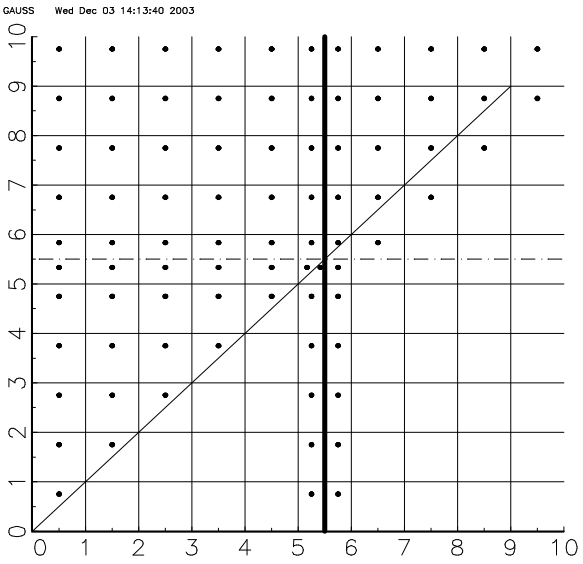


Figure 3: Illustration of Points of Support Necessary to Deal with Exclusion Restrictions

and the upper bound for $P(T_1 \leq t \,|X=1) - P(T_1 \leq t \,|X=0)$ is then the value of

$$\max \sum_{s_1 \leq t} \widetilde{p}(s_1, s_2) - \sum_{s_1 \leq t} p(s_1, s_2)$$

subject to the same constraints.

As in section 3.5, there is no guarantee that the sample analogs of these will be consistent estimators of the lower and upper bounds for $P(T_1 \leq t \,|X=1) - P(T_1 \leq t \,|X=0)$ as the sample analogs of the constraints may have no solution. To derive consistent estimators of these, first define $\widehat{f}$ by

$$\widehat{f} = \max \sum -v_i$$

subject to

$$v_k + \sum_{\substack{t_k < s_1 < t_k+1 \\ s_2 > s_1}} p(s_1, s_2) = \widehat{P}(T = t_k, I = 1 | X = 0),$$

$$v_{k+M} + \sum_{\substack{t_k < s_2 < t_k+1 \\ s_1 > s_2}} p(s_1, s_2) = \widehat{P}(T = t, I = 0 | X = 0),$$

$$v_{k+2M} + \sum_{\substack{t_k < s_1 < t_k+1 \\ s_2 > s_1}} \widetilde{p}(s_1, s_2) = \widehat{P}(T = t_k, I = 1 | X = 1),$$

$$v_{k+3M} + \sum_{\substack{t_k < s_2 < t_k+1 \\ s_1 > s_2}} \widetilde{p}(s_1, s_2) = \widehat{P}(T = t, I = 0 | X = 1),$$

$$v_{1+4M} + \sum_{s_1, s_2} p(s_1, s_2) = 1, \qquad v_{2+4M} + \sum_{s_1, s_2} \widetilde{p}(s_1, s_2) = 1,$$

$$\sum_{s_1} p(s_1,s_2) = \sum_{s_1} \widetilde{p}(s_1,s_2), p \qquad (s_1,s_2) \geq 0, \qquad \widetilde{p}(s_1,s_2) \geq 0$$

This has a feasible solution defined, for example, by setting $p(s_1,s_2) = \widetilde{p}(s_1,s_2) = 0$ for all (s_1,s_2).

The consistent estimator of the lower bound for $P(T_1 \leq t \,|X=1) - P(T_1 \leq t \,|X=0)$ is then the value of

$$\min \sum_{s_1 \leq t} \widetilde{p}(s_1,s_2) - \sum_{s_1 \leq t} p(s_1,s_2)$$

subject to

$$v_k + \sum_{\substack{t_k < s_1 < t_k+1 \\ s_2 > s_1}} p(s_1,s_2) = \widehat{P}(T = t_k, I = 1 | X = 0),$$

$$v_{k+M} + \sum_{\substack{t_k < s_2 < t_k+1 \\ s_1 > s_2}} p(s_1,s_2) = \widehat{P}(T = t, I = 0 | X = 0),$$

$$v_{k+2M} + \sum_{\substack{t_k < s_1 < t_k+1 \\ s_2 > s_1}} \widetilde{p}(s_1,s_2) = \widehat{P}(T = t_k, I = 1 | X = 1),$$

$$v_{k+3M} + \sum_{\substack{t_k < s_2 < t_k+1 \\ s_1 > s_2}} \widetilde{p}(s_1,s_2) = \widehat{P}(T = t, I = 0 | X = 1),$$

$$v_{1+4M} + \sum_{s_1,s_2} p(s_1,s_2) = 1, \qquad v_{2+4M} + \sum_{s_1,s_2} \widetilde{p}(s_1,s_2) = 1,$$

$$\sum_{s_1} p(s_1,s_2) = \sum_{s_1} \widetilde{p}(s_1,s_2)$$

$$\widehat{f} = \sum -v_i \qquad p(s_1,s_2) \geq 0, \qquad \widetilde{p}(s_1,s_2) \geq 0$$

3.7 Bounds with Continuous Covariates or Non–groups Durations

In the discussion above, we focused on the case where the explanatory variable X is discrete and the durations are grouped. This is the case in which the competing risk model with the parametric assumptions is most obviously not identified, and it therefore represents a worst–case scenario. On the other hand, it is also a case in which all the observed variables have a discrete distribution. This is essential for the simple approach taken above.

Each of the two complications, discrete covariates and grouped durations, violate the assumptions in for example Heckman and Honoré (1989) or Abbring and van den Berg (2003). In this section we demonstrate that it is in principle easy to derive expressions for the parameters of interest if only one of the two problems are present.

First assume that X is continuous and the durations are grouped. If the model is

$$(T^*, I) = (\min\{\alpha(X) S_1, \beta(X) S_2\}, 1\{\alpha(X) S_1 < \beta(X) S_2\})$$

with the normalization $\alpha(0) = \beta(0) = 1$, then the identified region for $(\alpha(\cdot), \beta(\cdot))$ is the set of functions $(a(\cdot), b(\cdot))$ such that there exists $p(s_1, s_2)$ satisfying

$$\int_{t_k/a(X)}^{t_{k+1}/a(X)} \int_{a(X)s_1/b(X)}^{\infty} p(s_1, s_2)\, ds_2 ds_1 = P(T = t_k, I = 1|\, X), \tag{14}$$

$$\int_{t_k/b(X)}^{t_{k+1}/b(X)} \int_{b(X)s_2/a(X)}^{\infty} p(s_1, s_2)\, ds_1 ds_2 = P(T = t_k, I = 0|\, X\ ), \tag{15}$$

$$\int\int p(s_1, s_2)\, ds_1 ds_2 = 1, \tag{16}$$

$$p(s_1, s_2) \geq 0 \tag{17}$$

for all values of X (where the first four equations hold for all $k = 1, .., M$). The identified region for $(\alpha(\cdot), \beta(\cdot), p(\cdot, \cdot))$ can also be expressed as

$$\arg \min_{a(\cdot), b(\cdot), p(\cdot,\cdot)} E\left[\sum_k \left(\int_{t_k/a(X)}^{t_{k+1}/a(X)} \int_{a(X)s_1/b(X)}^{\infty} p(s_1, s_2)\, ds_2 ds_1 g(X) - P(T = t_k, I = 1|\, X)\, g(X)\right)^2\right.$$
$$\left. + \sum_k \left(\int_{t_k/b(X)}^{t_{k+1}/b(X)} \int_{b(X)s_2/a(X)}^{\infty} p(s_1, s_2)\, ds_1 ds_2 g(X) - P(T = t_k, I = 0|\, X\ )\, g(X)\right)^2\right]$$

subject to $\int\int p(s_1, s_2)\, ds_1 ds_2 = 1$ and $p(s_1, s_2) \geq 0$ where $g(\cdot)$ is a positive weighting function. As discussed in Honoré and Tamer (2003), this can be turned into a feasible estimator of the identified region of (a, b) by replacing terms like $P(T = t_k, I = 1|\, X)$ by the nonparametric estimates and replacing a, b and p by approximations. The weighting function $g(\cdot)$ is useful because it can be used to control for the fact that $P(T = t_k, I = 1|\, X)$ will be imprecisely estimated in the tails of the distribution of X. In particular, it is straightforward to prove consistency of the estimator of the identified region for (α, β) if one uses $g(\cdot)$ to be the estimated density of X. Parametric restrictions on $\alpha(\cdot)$ and $\beta(\cdot)$ can be incorporated by minimizing the function above, subject to those restrictions.

Next consider the case where X is discrete with two points of support and the durations are not grouped. In this case, the identified region is given by the set of (a, b) for which there exists

$p(s_1, s_2)$ satisfying

$$\begin{aligned}
\int_t^\infty \int_{s_1}^\infty p(s_1, s_2)\, ds_2 ds_1 &= P(T > t, I = 1 | X = 0) \\
\int_t^\infty \int_t^{s_1} p(s_1, s_2)\, ds_2 ds_1 &= P(T > t, I = 0 | X = 0) \\
\int_{t/a}^\infty \int_{as_1/b}^\infty p(s_1, s_2)\, ds_2 ds_1 &= P(T > t, I = 1 | X = 1) \\
\int_{t/a}^\infty \int_{t/b}^{as_1/b} p(s_1, s_2)\, ds_2 ds_1 &= P(T > t, I = 0 | X = 1) \\
\int\int p(s_1, s_2)\, ds_1 ds_2 &= 1, \\
p(s_1, s_2) &\geq 0
\end{aligned}$$

This can also be expressed as the solutions to a population optimization problem,

$$\begin{aligned}
&\min_{a(\cdot), b(\cdot), p(\cdot,\cdot)} \int_0^\infty \left(\int_t^\infty \int_{s_1}^\infty p(s_1, s_2)\, ds_2 ds_1 - P(T > t, I = 1 | X = 0) \right)^2 dt \\
&+ \int_0^\infty \left(\int_t^\infty \int_t^{s_1} p(s_1, s_2)\, ds_2 ds_1 - P(T > t, I = 0 | X = 0) \right)^2 dt \\
&+ \int_0^\infty \left(\int_{t/a}^\infty \int_{as_1/b}^\infty p(s_1, s_2)\, ds_2 ds_1 - P(T > t, I = 1 | X = 1) \right)^2 dt \\
&+ \int_0^\infty \left(\int_{t/a}^\infty \int_{t/b}^{as_1/b} p(s_1, s_2)\, ds_2 ds_1 - P(T > t, I = 0 | X = 1) \right)^2 dt
\end{aligned}$$

subject to $\int\int p(s_1, s_2)\, ds_1 ds_2 = 1$ and $p(s_1, s_2) \geq 0$. This can be turned into a feasible estimator of the identified region of (a, b) by replacing terms like $P(T > t, I = 1 | X = 0)$ by the nonparametric estimates and replacing p by a sieve approximation.

4 The Change between 1970 and 1990 in the Mortality from Cancer and Cardiovascular Disease

In this section, we apply the methods described above to estimate the trends in disease-specific mortality between 1970 and 2000.

4.1 Data

We use mortality rates by single year of age, gender, race (black and white) and cause of death. These were calculated by matching population data from the Census Bureau and number of deaths from the Multiple Cause of Death Mortality files from 1970, 1980, 1990 and 2000. We computed

mortality rates for three causes of death. Deaths from cardiovascular diseases (hereafter CVD) included ICD8 and ICD9 codes 390-458, and ICD10 codes G45, G46 and I00-I99. Deaths from cancer included ICD8 and ICD9 codes 140-239, and ICD10 codes C00 through D48. Lung cancer includes ICD8 and ICD9 codes 162, and ICD10 code C34. All other diseases were counted under the category "other causes of death". We restrict the sample to individuals over age 45, so all the statistics we present are conditional on survival to that age. For 1970, population counts exist by single year of age up to age 79, and by 5-year intervals over age 80. To obtain consistent results over time, we therefore censor durations for all years at age 80.

4.1.1 Some Data Issues

There are several issues in calculating age-specific mortality rates using matched data from the census and the death certificate files.

Age misreporting both in the census and in death certificates are an important concern. To the extent that this error is not random, it may result in biased death rates. More importantly, these biases may have changed over time. In the census there is evidence of age heaping: individuals ages 50 and above tend to overstate their ages by "rounding up," which results in an unusually large population for ages ending in either 5 or 0. Figure 2 shows population counts by single year of age for 1980 (patterns for other years are similar): they show that in our data age heaping is mostly an issue for blacks. Another important issue (that cannot be fully separated from age misreporting) is that the census undercounts certain groups of the population, especially blacks, and the undercount varies with age. This problem is again larger for blacks than for whites. Furthermore, the extent of the undercount varies with the census year (Schenker (1993)). In the death certificates, there is also error in the age at death, but this error seems to be mostly confined to blacks over the ages of 65, who tend to understate their ages. There is no evidence of bias in ages among whites even for those above 85 (Hill, Preston, and Rosenwaike (2000)). Overall age misreporting appears to be a very important issue among blacks, although not much is known about how this may have changed over time. The overall effect of age misreporting is to downward-bias mortality for older cohorts (Preston, Elo, and Stewart (1999)). In the absence of additional data, there is no obvious way to correct mortality rates for these problems. So our results for blacks must be taken with caution.

Another issue is whether causes of death are correctly specified in the death certificate.[6] More importantly the issue is whether there have been significant changes from 1970 to 2000 in the

[6]For example Welch and Black (2002) report that deaths that follow surgery from cancer are not attributed to the cancer for which surgery was performed.

accuracy with which causes of death are reported. There were two changes in the International Classification of Diseases (ICD) during our period, one in 1978 (from ICD8 to ICD9) and another in 1998 (to ICD10). These changes have affected trends in mortality rates by cause, but previous research has suggested the effects of these classification changes are small for broad causes of death such as cancer and CVD (Jemal, Ward, Anderson, and Thun (2003), Klebba (1980) and Anderson, Minio, Hoyert, and Rosenberg (2001)). Furthermore, studies that have compared the causes of death reported in the death certificate with the cause of death from an autopsy, have found that the quality of death certificate reporting has not changed much since the 1960s, except perhaps for the very old (Hoel, Ron, Carter, and Mabuchi (1993)).

4.1.2 Descriptive Statistics

Table 1 presents summary statistics of the data (prior to censoring at age 80) for each census year and for four demographic groups defined by gender and race. It documents well-known patterns in mortality. As of 1970, between 55 and 70% of individuals died from CVD. However there were large differences across demographic groups in age at death from all causes and from cancer and CVD: White women lived the longest, followed by white men, black women and lastly black men. From 1970 to 2000, all groups experienced an increase in the age at death; and the share of individuals dying from cardiovascular disease fell dramatically while the share dying from cancer increased for all groups (although it fell in the 1990s for all except white men). But again there are some important differences across groups: the increase in life expectancy was largest for black females, the reductions in the percentage of CVD deaths were largest for whites and the percentage increases in deaths from cancer were largest for black men. Because of these differences we analyze the results separately for each group.

With our data we can calculate the observed hazard rates as follows

$$\lambda_j(t) = \frac{d\,P(T > t, J = j)/\,dt}{P(T > t)} \qquad \text{for} \qquad j = 1, 2 \tag{18}$$

Figures 4 and 5 show these sub–hazards for white males, white females, black males and black females for $\widetilde{T}_1$ and $\widetilde{T}_2$ for each census year. We present hazard rates for cancer and CVD separately. These hazard rates present in more detail the same trends that the summary statistics show. Hazard rates from CVD declined quite significantly in every decade for all groups. On the other hand, there is no discernible trend in cancer hazard rates. It is also clear that hazard rates are fairly different across demographic groups. From these graphs we also note that, as expected, hazard rates are much more volatile among blacks, especially at older ages. This is true for both causes of death, but it is more pronounced for cancer rates. Censoring at age 80 alleviates the problem somewhat

since hazard rates become even more volatile for older ages (not shown).

4.2 Results

We start by constructing bounds for the time trends under the assumption that the time dummy has a (different) multiplicative effect on the duration until death for both T_1 and T_2, but without assuming independence, as in section 3.3. We are interested in whether trends in cancer mortality show any improvement if we do not assume that cancer and CVD are independent risks. We do assume that the potential duration to other causes of death is independent of the potential duration until death from cancer (hereafter denoted by $\widetilde{T}_1$), and the potential duration until death from CVD (denoted by $\widetilde{T}_2$). This allows us to calculate the joint distribution of $\left(\widetilde{T}, \widetilde{J}\right) = \left(\min_j \left\{\widetilde{T}_j\right\}, \arg\min_j \left\{\widetilde{T}_j\right\}\right)$ for a hypothetical individual who has the 1970 (or 2000) hazard rate throughout his life. If the duration until death has not changed since 1970, then we will find bounds around one, i.e. the duration until death in 1970 will be identical to the duration until death in a later period. Bounds above one signal improvements.

We compute bounds for four demographic groups separately, and for three different periods, 1970 to 1980, 1970 to 1990, and 1970 to 2000. The results are in Table 2. For all groups we find that the CVD duration increased substantially, by about 40% for white males, 33% for blacks and 24% for white females. This increase was not concentrated in a single decade but was rather constant.

Age until death from cancer also increased for all groups during this period. This increase was about 10% for males and 15-20% for women, certainly smaller than the percentage increases for CVD, but not negligible. However for both black and white males the increase was mostly concentrated in the 1990s; from 1970 to 1990 the increases were small, about 3 to 6%. The same is not true for females, who saw some significant improvements in the 1970s and 1980s. Overall, these bounds support the idea that there was significant progress in cancer. Importantly note that all the bounds are tightly estimated (the range of the bounds is about 0.003 and the largest range is 0.028), and they never include one.

A caveat with these results is that they may be driven exclusively by changes in lung cancer rates. Lung cancer accounts for a large fraction of cancer deaths (about 50% for men and 10% for women) and it is mostly affected by smoking behavior throughout life. Deaths from lung cancer diminished in the 1990s because of decreases in smoking that started to take place in the 1960s and that are unrelated to progress in prevention and treatment since 1973 (Andersen, Remington, Trentham-Dietz, and Reeves (2002)). Therefore in Table 3 we present bounds for all cancers excluding lung cancer. This exclusion results in much larger improvements in cancer for all groups. In fact for women, once lung cancer is excluded, improvements in cancer are larger (in percentage

terms) than improvements in CVD. This result is consistent with the fact that age-adjusted lung cancer death rates increased for women through the period. Because lung cancer and CVD have a common risk, smoking, it may be incorrect to include lung cancer with the third cause of death which we treat as independent. We re-estimate non-lung cancer trends by grouping all other causes of death into the "other" category. Our results (Table 4) are very similar for whites, but very different for blacks: we no longer find any progress in cancer (in fact the bounds are below one for 2000) for black men; but we find even larger trends for black women. The lack of robustness for the results for blacks makes it difficult to make conclusions about the trends for this group.

Another limitation of the procedure we use is that it imposes a multiplicative effect of the time dummy on both cancer and CVD durations. Alternatively we estimate bounds for cancer that impose a multiplicative effect on cancer only (as in section 3.2.2). These results are presented in Table 5. In all cases, relaxing the parametric assumption for CVD results in bounds that are very large, typically ranging from about 0.5 to about 2.3. Furthermore, of the 12 bounds, only one set of bounds does not contain one (white females 1970-2000). It is therefore not possible to draw any conclusions from these results. Intuitively, this is not surprising: since CVD is the largest cause of death, imposing structure on its hazard improves estimation dramatically.

Aside from the substantive interest of the results, we want to assess the extent to which our methodology improves estimation, i.e. whether the results are affected by imposing independence (but imposing similar structural restrictions). We assess the issue in two ways.

First we estimate the parameters again, imposing independence. (See the appendix for a detailed description.) The results are in Table 6. We compare the coefficients on cancer and CVD with the bounds we presented in Tables 2 and 3. The coefficients for CVD are generally very similar with or without independence. On the other hand the coefficients for cancer are much smaller when we assume independence (whether or not lung cancer is included). Even more importantly, these coefficients are below one for black men and women. In other words, the assumption of independence would lead us to conclude that progress in cancer has either been small or that in fact cancer mortality has deteriorated.

Intuitively, imposing independence is unappealing if we think that the underlying processes have common determinants, for example if there are unobserved factors that affect both hazards. Whether or not this is true is case-specific and one may need out of sample information to determine it, such as medical knowledge. It is well known that smoking affects both CVD and lung cancer.[7] Therefore our methodology should result in drastically different results if we estimate bounds for

[7]See references in next section.

CVD and lung cancer. The results are in Table 7. We compare them with the results from Table 6. Again for CVD there is no change in the estimates, whereas the results are very different for lung cancer. Once we account for the (known) dependence between CVD and lung cancer, we find that there has been significant progress in lung cancer, and not just in the 1990s. Between 1970 and 2000 duration until lung cancer death increased by about 3% for black men, 17% for white men and 13% for white women. Without dependence we would have concluded that duration until death fell during this period for all groups, by as much as 45%. Only for black women do we still find a worsening of the duration once we account for dependence.

Interestingly, excluding or including only lung cancer has no (little) effect in our estimates of CVD progress. Neither does the imposition of independence, even though our results do suggest that cancer and CVD are dependent. Intuitively this is because CVD is the largest risk. One way to understand this result is to think of dependence as a form of sample selection. The potential for sample selection to generate bias depends not only on how different the excluded sample is, but also on how (relatively) large this group is. In this sense, the potential for sample selection bias is largest for the smallest risks. In practice, these results suggest that it may not be very important to consider dependence if one is interested in CVD, but it may be extremely important for all other risks, especially for the smaller ones.

4.3 Additional evidence

Our findings provide support for the hypothesis that cancer and CVD are in fact dependent risks (at the population level), and that there has been progress in both diseases, measured in terms of the increases in the underlying cause-specific survival rates. In this section we provide evidence from other sources consistent with these findings.

There are in fact several known common risks associated with both CVD and cancer. The American Heart Association lists smoking, drinking alcohol in large amounts, and obesity as factors that increase the likelihood of coronary heart disease, stroke, high blood pressure and hypertension. Moderate alcohol consumption and exercise on the other hand reduce blood pressure and coronary heart disease. The National Cancer Institute and the American Cancer Society also document that the same factors affect the risk of certain cancers. Smoking increases cancers of the respiratory system, as well as cancer of the bladder, pancreas, liver, uterus, kidney, stomach, colon and rectum, and some leukemias. Obesity increases the risk of endometrial cancer (cancer of the lining of the uterus), cancer of the colon, gall bladder, prostate, pancreas, kidney, esophagus, and postmenopausal breast cancers. Excessive alcohol increases the risk of cancer of the mouth, pharynx, larynx, esophagus, liver, and breast. Finally, moderate alcohol consumption and exercise are also

reduce cancer risk. Exercise is thought to reduce the risk of colon and breast cancers, independent of the impact of activity on weight. And moderate alcohol consumption may lower the risk of leukemia, skin, breast and prostate cancers. This evidence suggests that at the individual level, cancer and CVD are not independent risks.

Recall that even if risks are independent for individuals, risk heterogeneity in the population will result in dependent risks at the aggregate level. There is indeed substantial evidence of genetic differences among individuals with respect to their susceptibility to both CVD (Nabel (2003)) and cancer (e.g. Lynch and de la Chapelle (2003), Wooster and Weber (2003)).[8] Furthermore there are also large differences in the population in terms of exposure to environmental factors and behaviors that increase particular death risks. For example in 2000, high school dropouts were more than twice as likely to smoke than college educated individuals, women below poverty level were twice as likely as women in the highest income levels to be obese, married individuals were less likely to exercise than those that have never married, and Hispanics were less likely than non-Hispanics to drink (Schoenborn, Adams, Barnes, Vickerie, and Schiller (2004)). This evidence further supports our finding of positive dependence across risks.

Another important issue is whether there is any evidence that there were indeed innovations in terms of cancer prevention or treatment during the period we study, starting in the 1970s for women and mostly in the 1990s for men. We focus on improvements for the major cancer sites (excluding lung[9]), i.e. breast, prostate, colorectal and ovarian cancer. Survival from colorectal cancer, which disproportionately affects men, has improved because of a combination of earlier detection and improved treatment at earlier stages. Standard treatment for colorectal cancer changed in 1990, following a National Institutes of Health Conference recommendation, to include a combination of 5FU and leucovorin, two previously existing drugs (NIH Consensus Conference (1990)). Although treatment for prostate cancer remains controversial, clinical trials in the 1990s show promising effects of hormonal treatment (Howe, Wingo, Thun, Ries, Rosenberg, Feigal, and Edwards (2001)). Improvements to treat women's cancers started earlier. Mammographies started being routinely offered in the 1970s and studies in the 1970s and 1980s showed that early detection substantially improved mortality, especially for women above 50.[10] Breast cancer treatment started to change in the

[8]See web pages of the American Heart Association and the National Cancer Institute for additional cites.

[9]The fight against lung cancer has mostly focused on reducing tobacco consumption. This effort began with the Surgeon General Report in 1964 that first publicly announced that smoking increased the risk of lung cancer, and continues today. These efforts are reflected in the trends in lung cancer many years later. To our knowledge there is no evidence of other forms of progress in lung cancer. Since our trends cannot all be explained by smoking-related trends, we look at progress in other types of cancer.

[10]A review of the evidence by the U.S. Preventive Services Task Force is available at

1980s with the dissemination of adjuvant chemotherapy, including multi-agent chemotherapy and tamoxifen, and then additional changes were implemented in the early 1990s for postmenopausal women (Mariotto et al 2002). Treatment for ovarian cancer was modified in 1986 (NIH Consensus Conference (1995)) to include surgery and chemotherapy with a platinum compound (cisplatin or carboplatin) after publication of results from randomized trials which showed their effectiveness (Omura, Blessing, Ehrlich, Miller, Yordan, Creasman, and Homesley (1986)).

The trends that we estimate also reflect changes in lifestyle and demographic characteristics. Some of these may reflect improvements in knowledge about the factors associated with cancer and the dissemination of this information. In other words these trends may also include prevention. Furthermore these trends may include changes in demographics that are completely unrelated to scientific advances in cancer. Ultimately we cannot say with certainty that the trends we estimate are uniquely related to progress in treatment or whether they also reflect prevention and cohort effects.

5 Policy applications: Counterfactuals

We next use the results to answer two questions. First we ask what the contribution of cancer improvements to changes in mortality would have been in the absence of improvements for cardiovascular disease. In some sense, this is the measure by which cancer researchers would like to be judged.

Alternatively we ask what the changes in mortality would have been in the absence of improvements in cancer, given the changes in CVD. Under the assumption that the objective of funding R&D is to decrease mortality, this is the metric we want to use to calculate the benefits of cancer-specific funding.

To answer these questions, we will (some day) use the estimators described in section 3.5....

6 Conclusions and Limitations

In this paper we show that relatively weak parametric assumptions can dramatically improve identification in competing risks models. Using a semi parametric framework we estimate trends for cancer mortality without assuming that other risks are independent. We find that trends in cancer show much larger improvements than previously estimated. These improvements are not all due to changes in smoking for younger cohorts. Also not all improvements took place in the 1990s; for

http://www.ahrq.gov/clinic/3rduspstf/breastcancer/brcanrr.htm#ref4

women, we find significant improvements going back to the 1970s.

7 Bibliography

References

Abbring, J. H., and G. J. van den Berg (2003): "The identifiability of the mixed proportional hazards competing risks model," *Journal of the Royal Statistical Society B*, 65, 701–710.

Ahn, H., and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58(1-2), 3–29.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

Andersen, L. D., P. Remington, A. Trentham-Dietz, and M. Reeves (2002): "Assessing a Decade of Progress in Cancer Control," *The Oncologist*, 7, 200–204.

Anderson, R. N., A. M. Minio, D. L. Hoyert, and H. M. Rosenberg (2001): "Comparability of Causes of Death Between ICD-9 and ICD-10: Preliminary Estimates," *National Vital Statistics Reports*, 49(2).

Bailar, J. C., and H. L. Gornik (1997): "Cancer Undefeated," *The New England Journal of Medicine*, 336.

Bailar, J. C., and B. M. Smith (1986): "Progress against cancer?," *The New England Journal of Medicine*, 314, 1226–1232.

Berrington, A., and I. Diamond (2000): "Marriage or Cohabitation: A competing risks analysis of the first-partnership formation among the 1958 birth cohort," *Journal of the Royal Statistical Society, Series A*, 163, 127–152.

Booth, A. L., and S. E. Satchell (1995): "The Hazards of doing a PhD: An Analysis of Completion and Withdrawal Rates of British PhD Students in the 1980s," *Journal of the Royal Statistical Society, Series A*, 158(2), 297–318.

Cox, D. R. (1962): *Renewal Theory*, Muthuen's Monographs on Applied Probability and Statistics. Methuen and Co., London.

Crowder, M. (2001): *Classical Competing Risks*. Chapman and Hall/CRC.

Deng, Y., J. M. Quigley, and R. Van Order (2000): "Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options," *Econometrica*, 68(2), 275–307.

Doll, R. (1991): "Progress against cancer: An epidemiologic assessment," *American Journal of Epidemiology*, 134(7).

Faraggi, D., and E. L. Korn (1996): "Competing Risks with Frailty Models When Treatment Affects Only One Failure Type," *Biometrika*, 83(2), 467–471.

Flinn, C. J., and J. J. Heckman (1982): "New Methods for Analyzing Structural Models of Labor Force Dynamics," *Journal of Econometrics*, 18, 115–168.

Heckman, J. J., and B. E. Honoré (1989): "The Identifiability of the Competing Risks Model," *Biometrika*, 76, 325–330.

——— (1990): "The Empirical Content of The Roy Model," *Econometrica*, 58, 1121–1149.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535.

Hill, M. E., S. H. Preston, and I. Rosenwaike (2000): "Age Reporting Among White Americans Aged 85+: Results Of A Record Linkage Study," *Demography*, 37(2).

Hoel, D. G., E. Ron, R. Carter, and K. Mabuchi (1993): "Influence of Death Certificate Errors on Cancer Mortality Trends," *Journal of the National Cancer Institute*, 85(13), 1063–8.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, 533–565.

Honoré, B. E., and E. T. Tamer (2003): "Bounds on Parameters in Dynamic Discrete Choice Models," Princeton University.

Howe, H., P. Wingo, M. J. Thun, L. A. Ries, H. M. Rosenberg, E. G. Feigal, and B. K. Edwards (2001): "Annual Report to the Nation on the Status of Cancer (1973 through 1998), Featuring Cancers with Recent Increasing Trends," *Journal of the National Cancer Institute*, 93(11).

Jemal, A., E. Ward, R. N. Anderson, and M. J. Thun (2003): "Influence of Rules From the Tenth Revision of the International Classification of Diseases on U.S. Cancer Mortality Trends," *Journal of the National Cancer Institute*, 95, 1727–1728.

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

Katz, L. F., and B. D. Meyer (1990): "Unemployment Insurance, Recall Expectations, and Unemployment Outcomes," *Quarterly Journal of Economics*, 105, 973–1002.

Klebba, A. J. (1980): "Estimates of Selected Comparability Ratios Based on Dual Coding of the 1976 Death Certificates by the Eighth and Ninth Revision of the International Classification of Diseases," *Monthly Vital Statistics Report*, 28(11).

Lin, D. Y., J. M. Robins, and L. J. Wei (1996): "Comparing Two Failure Time Distributions in the Presence of Dependent Censoring," *Biometrika*, 83(2), 381–393.

Lynch, H., and A. de la Chapelle (2003): "Hereditary colorectal cancer," *New England Journal of Medicine*, 348, 919–932.

Manski, C. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80, 319–323.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*. Springer Series in Statistics.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

Meyer, B. D. (1990): "Unemployment Insurance and Unemployment Spells," *Econometrica*, 58, 757–782.

Nabel, E. G. (2003): "Genomic Medicine: Cardiovascular Disease," *New England Journal of Medicine*, 349, 60–72.

NIH Consensus Conference, . (1990): "Adjuvant Therapy for Patients with Colon and Rectal Cancer," *Journal of the American Medical Association*, 264, 1444–1450.

——— (1995): "Ovarian Cancer, Screening, Treatment and Followup," *Journal of the American Medical Association*, 273, 491–497.

Omura, G., J. A. Blessing, C. E. Ehrlich, A. Miller, E. Yordan, W. T. Creasman, and H. D. Homesley (1986): "A Randomized Trial of Cyclophosphamide abd Doxorubicin with or Without Cisplatin in Advanced Ovarian Carcinoma: A Gynecologic Oncology Group Study," *Cancer*, 57, 1725–1730.

Peterson, A. V. (1976): "Bounds for a joint distribution with fixed sub–distribution functions: Application to competing risks," *Proceedings of the National Academy of Science*, 73, 11–13.

Prentice, R. L., and L. A. Gloeckler (1978): "Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data," *Biometrics*, 34, 57–67.

Preston, S. H., I. T. Elo, and Q. Stewart (1999): "Effects of Age Misreporting on Mortality Estiamtes at Older Ages," *Population Studies*, 53(2), 165–177.

Roy, A. D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers (New Series)*, 3, 135–146.

Schenker, N. (1993): "Undercount in the 1990 Census: Special Section," *Journal of the American Statistical Association*, 88(423).

Schoenborn, C., P. F. Adams, P. M. Barnes, J. L. Vickerie, and J. S. Schiller (2004): "Health Behaviors of Adults: United States, 1999–2001," *National Center for Health Statistics. Vital Health Stat*, 10(219).

Slud, E., and D. Byar (1988): "How Dependent Causes of Death Can Make Risk Factors Appear Protective," *Biometrics*, 44(1), 265–269.

Tsiatis, A. (1975): "A Nonidentifiability Aspect of the Problem of Competing Risks," *Proceedings of the National Academy of Sciences*, 72, 20–22.

Vaupel, J. W., and A. I. Yashin (1999): "Cancer Rates over Age, Time and Place: insights from Stochastic Models of Heterogrnous Populations," *Max Plank Institute for Demographic Reseach Working Paper 1999-006*.

Welch, H. G., and W. C. Black (2002): "Are Deaths Within a Month of Cancer-Directed Surgery Attributed to Cancer?," *Journal of the National Cancer Institute*, 94(14).

Welch, H. G., L. M. Schwartz, and S. Woloshin (2000): "Are increasing 5-year survival rates evidence of success against cancer?," *Journal of the American Medical Association*, 283(22), 2975–2978.

Wooster, R., and B. L. Weber (2003): "Genomic Medicine: Breast and Ovarian Cancer," *New England Journal of Medicine*, 348, 2339–2347.

8 Appendix

8.1 The Data

Population data

These data come from April 1st population counts from the Census Bureau, from the following sources:

1. 1970 population counts obtained from U.S. Bureau of the Census, Census of Population: 1970 General Population Characteristics Final Report PC(1)-B1 United States Summary.

2. 1980 Data was found at

   http://www.census.gov/population/estimates/nation/e80s/E8081RQI.txt

3. 1990 data was found at

   http://www.census.gov/population/estimates/nation/e90s/E9090RMP.txt

4. 2000 White population counts obtained from Census table PCT12A, Black population counts from table PCT12B and total population counts from PCT12. All three tables were found at the US Census Bureau's website: http://factfinder.census.gov/servlet

8.2 Calculating Coefficients under Independence

One can calculate the marginal distributions of T_1 and T_2 up to the censoring point under the assumption that the risks are independent. Focussing on the same risk in two different years and defining $Y = \max\{\log(c) - \log(T), 0\}$ where c is the censoring point, one gets the distribution of two random variables Y_1 and Y_2 satisfying

$$Y_1 = \max\{\varepsilon_1, 0\}$$

and

$$Y_2 = \max\{\varepsilon_2 - k, 0\}$$

where ε_1 and ε_2 are identically distributed and k is the log of the multiplicative constant.

These can also be written as

$$Y_1 = \max\{v_1 + k, 0\}$$

and

$$Y_2 = \max\{v_2, 0\}$$

There are then a number of ways to back out k. Here we follow the approach in Honoré (1992) and use the moment condition

$$\begin{aligned} & E\left[1\left\{Y_2 > \max\left\{0, -k, Y_1 - k\right\}\right\} - 1\left\{Y_1 > \max\{0, k, Y_2 + k\}\right]\right. \\ = \; & E\left[1\left\{\max\left\{v_2, 0\right\} > \max\left\{0, -k, \max\left\{v_1 + k, 0\right\} - k\right\}\right\}\right. \\ & \left. - 1\left\{\max\left\{v_1 + k, 0\right\} > \max\{0, k, \max\left\{v_2, 0\right\} + k\right\}\right] \\ = \; & E\left[1\left\{v_2 > \max 0, -k, v_1\right\} - 1\left\{v_1 + k > \max\{0, k, v_2 + k\right\}\right] \\ = \; & E\left[1\left\{v_2 > \max 0, -k, v_1\right\} - 1\left\{v_1 > \max\{-k, 0, v_2\right\}\right] \end{aligned}$$

If (v_1, v_2) is distributed like (v_2, v_1) then the last expression is 0 at the true k. Moreover, the first expression is monotone in k. This allows one to solve for a unique k.

8.3 Details about the Calculations

The function value that defines the identified region was calculated over three grids.

The first grid was defined by the rectangle $\{0.90, 0.95, 1.00, ..., 1.40\} \times \{0.90, 0.95, 1.00, ..., 1.40\}$.

The second grid was defined by first calculation the set of maximizers over the original grid. Let $\theta_1^{\min}$ and $\theta_1^{\max}$ denote the minimum and maximum value of the first coordinate in that set and let $\theta_2^{\min}$ and $\theta_2^{\max}$ denote the minimum and maximum value of the second coordinate in the set. The second grid is then given by $\left\{\theta_1^{\min} - 0.05, \theta_1^{\min} - 0.04, \theta_1^{\min} - 0.03, ..., \theta_1^{\max} + 0.08\right\} \times \left\{\theta_2^{\min} - 0.05, \theta_2^{\min} - 0.04, \theta_2^{\min} - 0.03, ..., \theta_2^{\max} + 0.08\right\}$.

The third grid was defined in terms of the maximizers over the first two grid. Let $\theta_1^{\min}$ and $\theta_1^{\max}$ denote the minimum and maximum value of the first coordinate in that set and let $\theta_2^{\min}$ and $\theta_2^{\max}$ denote the minimum and maximum value of the second coordinate in the set. The second grid is then given by $\left\{\theta_1^{\min} - 0.01, \theta_1^{\min} - 0.009, \theta_1^{\min} - 0.008, ..., \theta_1^{\max} + 0.015\right\} \times \left\{\theta_2^{\min} - 0.01, \theta_2^{\min} - 0.009, \theta_2^{\min} - 0.008, ..., \theta\right.$

The estimated identified region is then the set of maximizers of the union of the three grids. The numbers reported in the tables are the minimum and maximum values of each coordinate.

TABLE 1: Summary statistics by race, gender and decade (conditional on survival to age 45)

	1970	1980	1990	2000
White Males				
Age at death–all causes	70.43	72.0	73.62	74.70
Age at death from cardiovascular disease	71.57	72.99	74.51	75.97
Age at death from cancer	69.12	70.40	71.75	72.67
Age at death from other causes	68.18	70.96	73.32	74.17
Fraction deaths from cardiovascular disease	0.63	0.58	0.50	0.44
Fraction deaths from cancer	0.14	0.17	0.19	0.20
White Females				
Age at death–all causes	74.65	76.89	78.8	80.2
Age at death from cardiovascular disease	77.31	79.50	81.24	82.77
Age at death from cancer	68.37	70.54	72.57	73.86
Age at death from other causes	71.76	75.38	78.86	80.14
Fraction deaths from cardiovascular disease	0.62	0.59	0.51	0.45
Fraction deaths from cancer	0.17	0.19	0.19	0.18
Black Males				
Age at death–all causes	66.09	68.09	69.4	69.23
Age at death from cardiovascular disease	67.65	69.50	70.43	70.44
Age at death from cancer	66.30	67.90	69.42	69.73
Age at death from other causes	63.10	65.85	67.76	67.54
Fraction deaths from cardiovascular disease	0.56	0.51	0.46	0.43
Fraction deaths from cancer	0.14	0.18	0.21	0.21
Black Females				
Age at death–all causes	68.21	71.42	73.64	74.74
Age at death from cardiovascular disease	70.18	73.46	75.47	76.87
Age at death from cancer	64.63	67.30	69.39	70.21
Age at death from other causes	65.50	69.86	73.35	74.34
Fraction deaths from cardiovascular disease	0.61	0.56	0.51	0.46
Fraction deaths from cancer	0.15	0.18	0.20	0.19

TABLE 2: Marginal Identified Regions

Results for White Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.126, 1.129)$	$(1.295, 1.296)$	$(1.389, 1.391)$
Coefficient on Cancer	$(1.001, 1.029)$	$(1.020, 1.035)$	$(1.134, 1.153)$

Results for White Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.092, 1.093)$	$(1.160, 1.160)$	$(1.236, 1.238)$
Coefficient on Cancer	$(1.091, 1.092)$	$(1.154, 1.157)$	$(1.201, 1.206)$

Results for Black Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.126, 1.129)$	$(1.201, 1.206)$	$(1.334, 1.346)$
Coefficient on Cancer	$(1.030, 1.034)$	$(1.063, 1.066)$	$(1.072, 1.074)$

Results for Black Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.158, 1.159)$	$(1.231, 1.235)$	$(1.334, 1.346)$
Coefficient on Cancer	$(1.096, 1.096)$	$(1.167, 1.172)$	$(1.158, 1.159)$

TABLE 3: Marginal Identified Regions

Results for White Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.126, 1.129)	(1.295, 1.296)	(1.392, 1.399)
Coef. on Cancer (excl. lung)	(1.091, 1.093)	(1.039, 1.045)	(1.236, 1.249)

Results for White Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.091, 1.093)	(1.201, 1.206)	(1.267, 1.269)
Coef. on Cancer (excl. lung)	(1.126, 1.129)	(1.239, 1.249)	(1.455, 1.458)

Results for Black Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.126, 1.129)	(1.202, 1.206)	(1.334, 1.346)
Coef. on Cancer (excl. lung)	(1.112, 1.115)	(1.201, 1.205)	(1.118, 1.119)

Results for Black Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.154, 1.157)	(1.286, 1.296)	(1.334, 1.346)
Coef. on Cancer (excl. lung)	(1.106, 1.111)	(1.143, 1.148)	(1.308, 1.319)

TABLE 4: Marginal Identified Regions

Results for White Males

	1970–1980	1970–1990	1970–2000
Coefficient on All Other	$(1.091, 1.093)$	$(1.201, 1.206)$	$(1.239, 1.249)$
Coef. on Cancer (excl. lung)	$(1.223, 1.230)$	$(1.001, 1.062)$	$(1.191, 1.195)$

Results for White Females

	1970–1980	1970–1990	1970–2000
Coefficient on All Other	$(1.091, 1.092)$	$(1.101, 1.103)$	$(1.112, 1.115)$
Coef. on Cancer (excl. lung)	$(1.092, 1.093)$	$(1.334, 1.346)$	$(1.467, 1.473)$

Results for Black Males

	1970–1980	1970–1990	1970–2000
Coefficient on All Other	$(1.091, 1.093)$	$(1.154, 1.159)$	$(1.236, 1.249)$
Coef. on Cancer (excl. lung)	$(1.126, 1.129)$	$(1.001, 1.060)$	$(0.990, 0.999)$

Results for Black Females

	1970–1980	1970–1990	1970–2000
Coefficient on All Other	$(1.126, 1.129)$	$(1.201, 1.206)$	$(1.191, 1.199)$
Coef. on Cancer (excl. lung)	$(1.094, 1.126)$	$(1.158, 1.159)$	$(1.450, 1.458)$

TABLE 5: Marginal Identified Regions
(only Cancer multiplicative)

Results for White Males

	1970–1980	1970–1990	1970–2000
Coefficient on Cancer	$(0.520, 2.186)$	$(0.602, 2.124)$	$(0.654, 2.124)$

Results for White Females

	1970–1980	1970–1990	1970–2000
Coefficient on Cancer	$(0.802, 1.610)$	$(0.890, 1.646)$	$(1.002, 1.698)$

Results for Black Males

	1970–1980	1970–1990	1970–2000
Coefficient on Cancer	$(0.449, 2.356)$	$(0.484, 2.200)$	$(0.550, 2.332)$

Results for Black Females

	1970–1980	1970–1990	1970–2000
Coefficient on Cancer	$(0.556, 2.284)$	$(0.644, 2.230)$	$(0.702, 2.332)$

TABLE 6: Results Assuming Independence

Results for White Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	1.125	1.250	1.385
Coefficient on Cancer	1.029	1.029	1.059
Coef. on Cancer (excl. lung)	1.091	1.091	1.091
Coef. on Lung Cancer	0.889	0.889	0.944

Results for White Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	1.091	1.200	1.263
Coefficient on Cancer	1.034	1.059	1.091
Coef. on Cancer (excl. lung)	1.091	1.200	1.333
Coef. on Lung Cancer	0.639	0.556	0.556

Results for Black Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	1.125	1.222	1.333
Coefficient on Cancer	0.972	0.941	0.972
Coef. on Cancer (excl. lung)	1.059	1.091	1.091
Coef. on Lung Cancer	0.778	0.722	0.833

Results for Black Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	1.161	1.286	1.333
Coefficient on Cancer	0.972	0.972	1.029
Coef. on Cancer (excl. lung)	1.059	1.125	1.200
Coef. on Lung Cancer	0.583	0.528	0.556

TABLE 7: Marginal Identified Regions

Results for White Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.126, 1.129)$	$(1.239, 1.249)$	$(1.389, 1.391)$
Coef. on Lung Cancer	$(0.962, 0.962)$	$(1.072, 1.103)$	$(1.179, 1.181)$

Results for White Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.084, 1.086)$	$(1.154, 1.157)$	$(1.201, 1.206)$
Coef. on Lung Cancer	$(1.039, 1.039)$	$(1.001, 1.029)$	$(1.134, 1.136)$

Results for Black Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.126, 1.129)$	$(1.154, 1.166)$	$(1.334, 1.346)$
Coef. on Lung Cancer	$(0.929, 0.931)$	$(1.084, 1.153)$	$(1.032, 1.032)$

Results for Black Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	$(1.154, 1.157)$	$(1.231, 1.235)$	$(1.334, 1.346)$
Coef. on Lung Cancer	$(0.879, 0.886)$	$(1.160, 1.166)$	$(0.945, 0.959)$

Figure 1a: Trends in Age-Adjusted Mortality 1950-2000 (all persons)

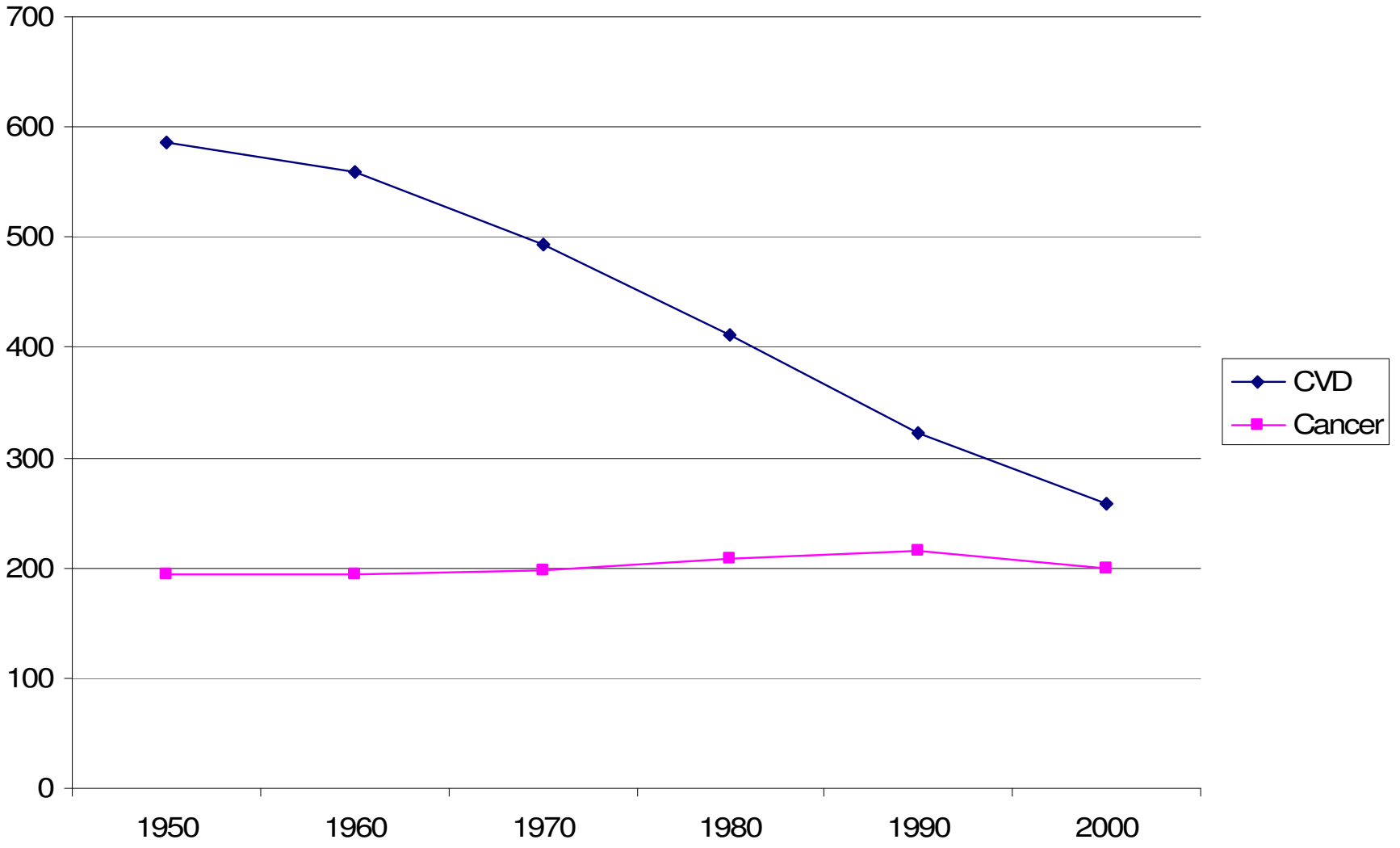
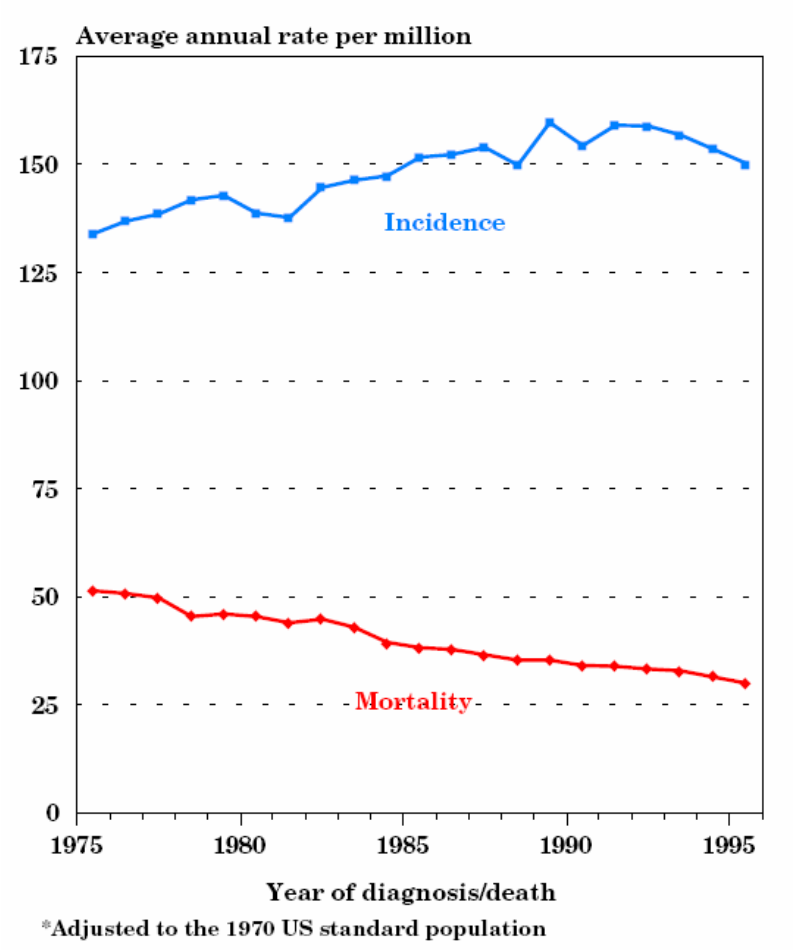
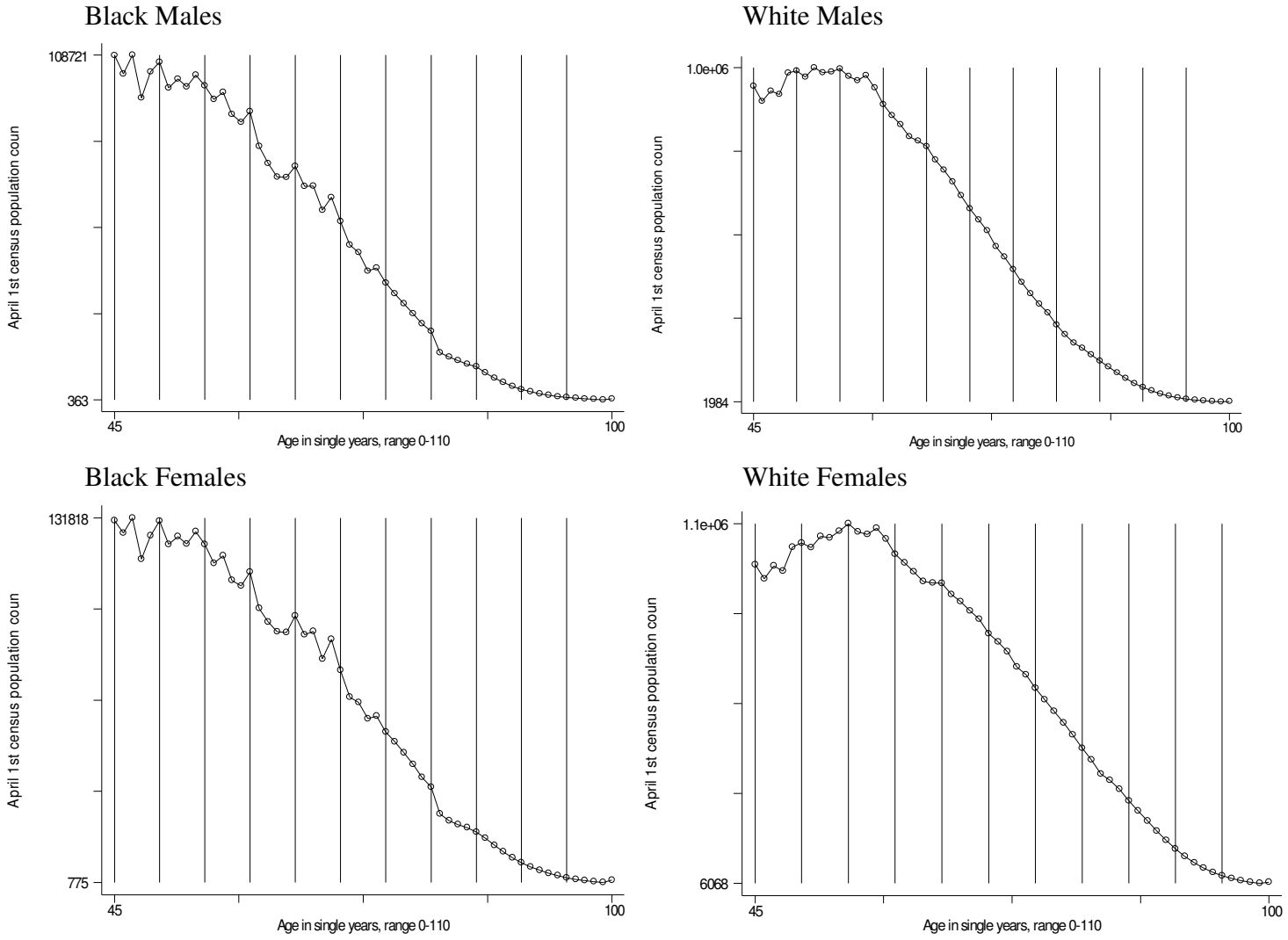


Figure 1b: Trends in incidence and age-adjusted mortality for individuals ages 20 and below



Source: Ries LAG, Smith MA, Gurney JG, Linet M, Tamra T, Young JL, Bunin GR (eds). Cancer Incidence and Survival among Children and Adolescents: United States SEER Program SEER Program 1975-1995, National Cancer Institute, SEER Program. NIH Pub. No. 99-4649. Bethesda, MD, 1999.

Figure 2: 1980 April 1st population counts by gender and race, ages 45 and above



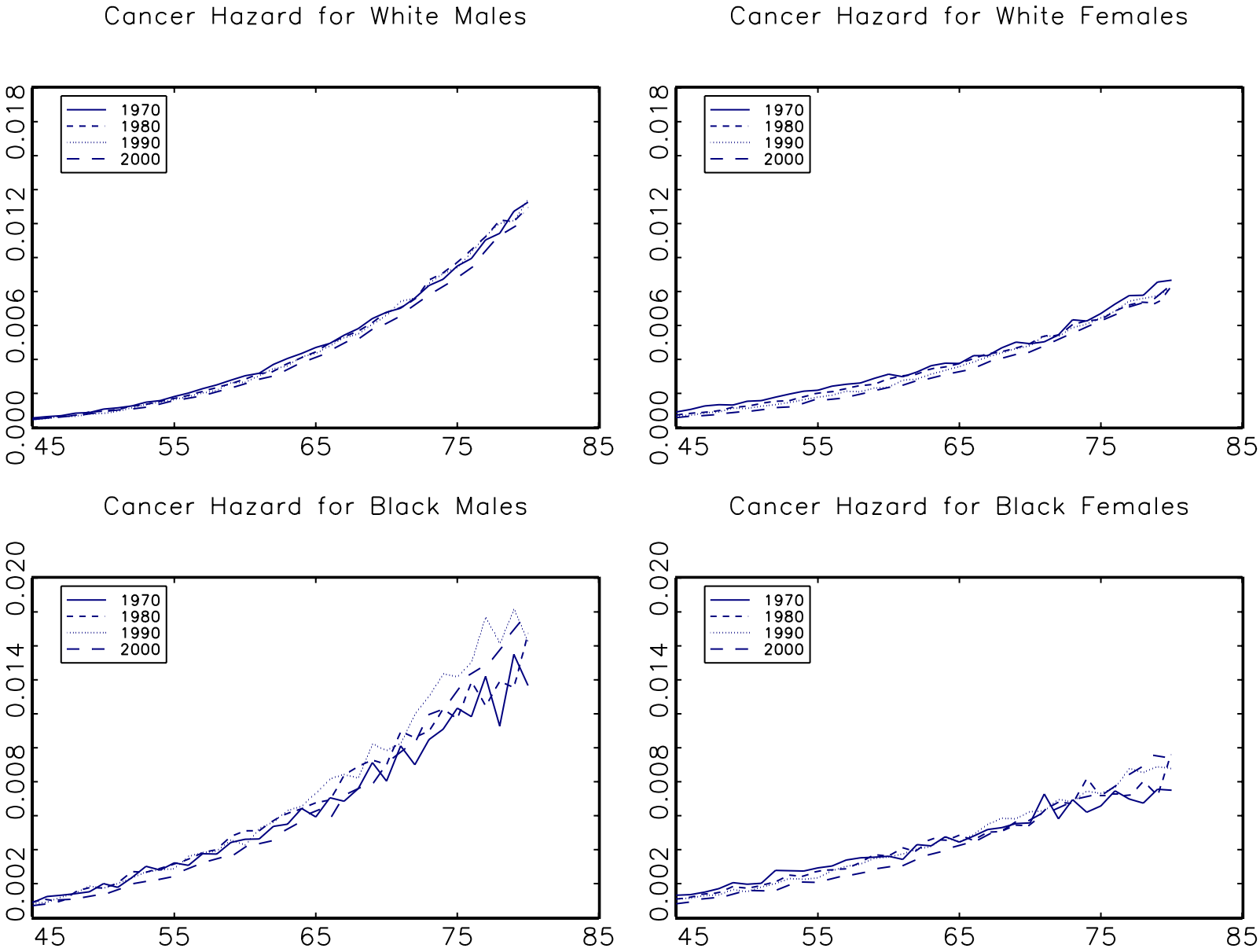


Figure 4: Hazard Rates for the Cancer

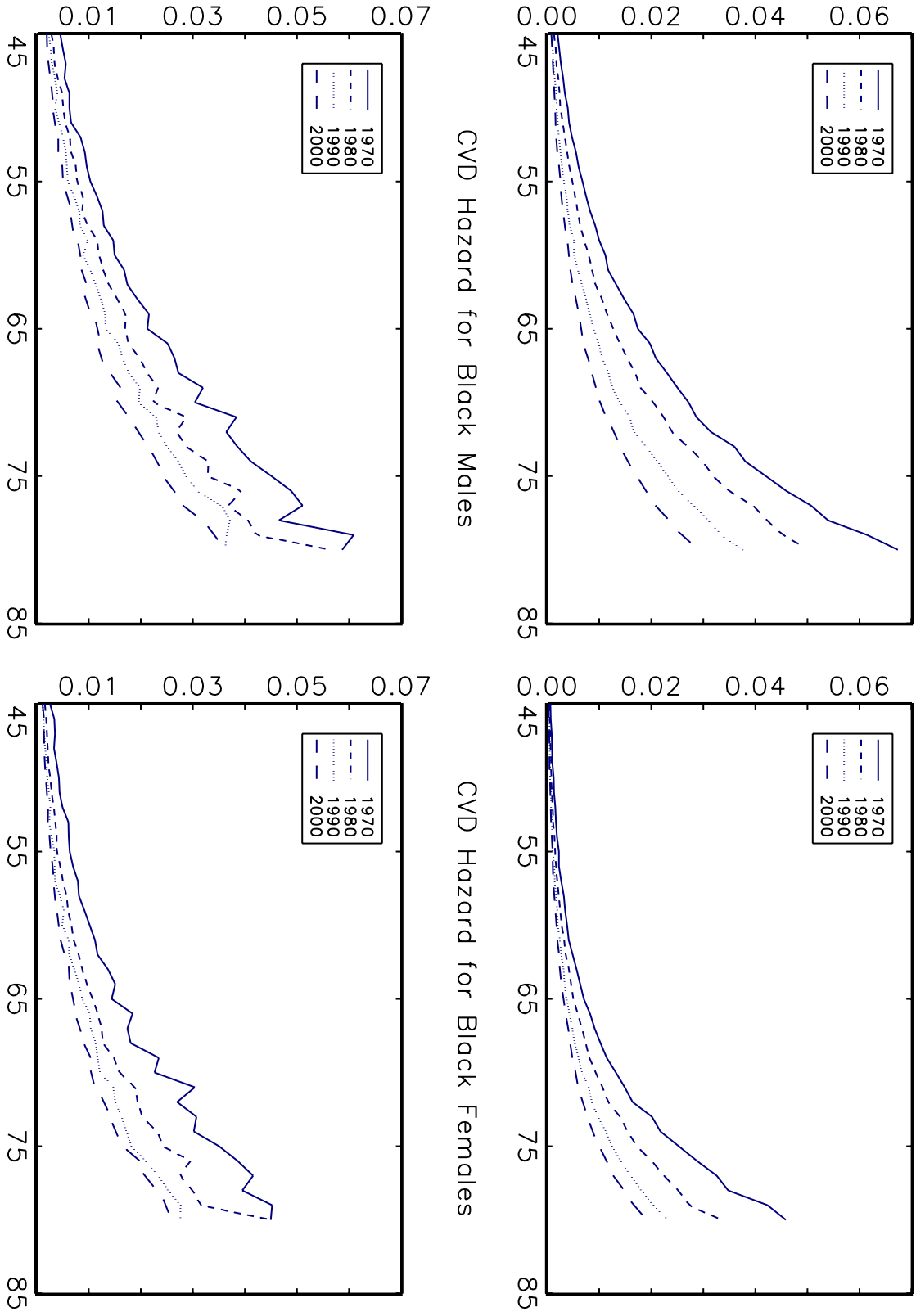


Figure 5: Hazard Rates for the CVD