Bounds in Competing Risks Models and the War on Cancer*

Bo E. Honoré† Adriana Lleras-Muney‡

This Version: June 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.

Keywords: Bounds, Competing Risks, Cancer, Cardiovascular.

JEL Classification: I10, C40

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 confirm this view: age-adjusted incidence rates and mortality rates show a bleak picture. Age-adjusted mortality from cancer increased from 198.7 (per 100,000) in 1973 to 213 in 1993, and

*We would like to thank Jaap Abbring, Josh Angrist, Eric J. Feuer, researchers at the National Cancer Institute, and seminar participants at CAM at the University of Copenhagen, L. S. E. , M. I. T., U. C. L. and the Harvard–M. I. T.–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).

then it fell to about its 1973 levels (198.6) in 2000. Incidence rates show a similar pattern increasing from 385 in 1973 to 509.85 in 1992, and then decreasing to 477 in 2000 (SEER (2004)).

At the same time, age-adjusted mortality rates from cardiovascular disease have fallen quite dramatically. (See Figure 3.) 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. Moreover these reductions have occurred in spite of the increases in cancer incidence for both groups (Doll (1991)). 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.[2] 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 the underlying risks are independent, 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. 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,

[2]Chiang (1991), Rothenberg (1994) and Llorca and Delgado-Rodriguez (2001) have investigated the effects of cardiovascular mortality trends on trends in cancer mortality. However Wohlfart and Andersen (2001) point out, these authors assume that risks are independent in their analysis.

Smith, and Clements (1997).

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

2 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: cardiovascular disease (hereafter CVD), cancer and all other causes. (For data sources and details see the appendix.) We restrict the sample to individuals over age 45, so all the results 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.

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 the 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

[3]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 (Welch, Schwartz, and Woloshin (2000)). Similarly improved detection increases observed incidence, even though disease rates may not have changed. 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.

we analyze the results separately for each group.

With our data we can calculate the observed hazard rates. Figures 5 and 6 show these sub–hazards for white males, white females, black males and black females 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 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).

3 Competing Risks

In this section we review the theory on competing risks, illustrating issues and methods in the context of cancer and cardiovascular mortality and using the data we just described.

3.1 Set–up

Formally, a competing risks model is a duration model where the observed duration is the shortest of a number of latent durations. In addition it is typically also assumed that the identity of the shortest duration is observed. Mathematically, we observe T and δ where

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

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 (latent) duration until death from a specific cause (risk) such as cancer or cardiovascular disease, T the observed duration until death and δ the cause of death. In order to simplify the exposition and to present the theory related to the specific case we analyze, 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.

In this paper we will use the notation

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

and the object of interest will be features of the distribution of (T_1, T_2) given a set of explanatory variables X. Knowledge of the joint distribution of the unobserved, latent distributions T_1 and

T_2 (given X) allows one to answer policy questions that one could not answer on the basis of the distribution of (T^*, δ) (given X). For example, the latter will not allow one to evaluate the effect of eliminating one of the risks on the distribution of the duration until death.

As discussed below, applications of competing risks models have often, though not always, assumed that the underlying latent durations are statistically independent. While such an assumption is reasonable in some contexts, there are at least two related reasons why one could suspect it to be violated in specific situations.

The first reason why the latent durations might be dependent is that the same underlying process affects both risks. In the case of CVD and cancer, there are several common risk factors that affect both. 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 other cancers. Obesity increases the risk of cancer of the uterus, breast and prostate cancer among others. Excessive alcohol increases the risk of cancer of the mouth, pharynx, larynx, esophagus, liver, and breast. Exercise is thought to reduce the risk of colon and breast cancers, 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.

Additionally, heterogeneity across individuals can cause the underlying latent durations to be dependent even if the risks are independent for every individual in that population (Vaupel and Yashin (1999)). There is substantial evidence of genetic differences across 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)).[4] This will lead the the latent duration until death from CVD and cancer to be correlated. 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)).

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

This suggests that it is interesting to consider competing risks models with dependent latent durations.

3.2 Identification

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 $(\min\{T_1, T_2\}, 1\{T_1 < T_2\})$ equals that of $(\min\{S_1, S_2\}, 1\{S_1 < S_2\})$ (see Cox (1962) and Tsiatis (1975)). In other words, for every dependent distribution of (T_1, T_2), one can find an independent distribution that generates observationally equivalent data. Since this exercise can be carried out conditional on a set of explanatory variables X, the relationship between T_1 and T_2 conditionally on X is fundamentally unidentified, and it is not possible to use observational data only to test whether or not the risks are dependent. It is therefore necessary to make additional assumptions if one wants to answer questions that require exact knowledge of the joint distribution of (T_1, T_2).

Broadly speaking, there have been three approaches to dealing with the identification problem in competing risks. The first is to make no additional assumptions and to estimate bounds for the object of interest, for example the marginal distributions of the underlying durations. The second approach is to assume that the risks are independent (conditional on a set of observed covariates) in which case estimation of competing risks models amounts to estimation of duration models with random censoring. The third broad approach is to specify a parametric or semiparametric model for (T_1, T_2) conditional of the covariates. The approach taken in this paper is a combination of the first and the third approach.

If one is willing to assume independence then it is straightforward to estimate the hazard function for each of the underlying distributions. For the case of cancer, the hazard rates in Figure 5 are sufficient to conclude that there has been a very small improvement in cancer mortality, if any at all. Of course, imposing independence when the risks are indeed dependent, will result in inconsistent estimates of the cause-specific hazard rates and of the effect of covariates on those hazards.[5] Given that the medical evidence suggests that CVD and cancer are dependent, it is therefore not possible to reach definite conclusions by looking at the observed hazards, as we did above.

[5]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. Slud and Byar (1988) provide such an example. Vaupel and Yashin (1999) illustrate the problems that arise if one assumes independence in the presence of unobserved population heterogeneity (which results in dependent population hazards).

Alternatively, one can make no assumptions on the joint distribution of the survival rates, and estimate bounds on the objects of interest. Following the approach of, for example, Peterson (1976) and Manski (2003), it is straightforward to generate bounds on the marginal distributions of T_1 and T_2. These bounds are given in Peterson (1976), who also provides bounds on the joint distribution of T_1 and T_2. It is easy to understand the basic idea behind these bounds. For example suppose that 15% of individuals have died of CVD by age 60 and 10% have died of cancer. The survival rate from cancer at age 60 can be bounded between 75 and 90%. Although this approach is very appealing, the nonparametric bounds are generally very wide (see the numerical example in Peterson (1976)), making it difficult to draw conclusions. In Figure 4 we present the bounds for the survival from cancer in 1970 and 2000 for our four demographic groups. It is evident from these graphs that it is not possible to make any statement about whether survival from cancer has increased or decreased in this time period.

The results presented in Figure 4 and the potential problems with assuming independence, suggest 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. At the extreme, one could specify a fully parametric model and estimate the parameters of such a model by maximum likelihood. This is the approach taken in most of the applications cited in the introduction. The weakness of a fully parametric approach is that one worries about the extent to which the results are driven entirely by the functional form assumptions. A number of papers have therefore studied identifiability of semiparametric competing risks models.

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 $\Re^2_+$. 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, as discussed by Crowder (2001) the conditions for identification are restrictive and often unrealistic as the covariates of interest have bounded support and are not continuous in many applications. For example, analyzes 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. However, the durations are observed in groups in many data sets. 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). This is the subject of the next section.

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.

4 Bounds in Some Specific Competing Risks Models

As mentioned above, one of the motivations for this paper is that many data sources contain interval observations on durations, whereas the results on identification of semiparametric competing risks models assume that durations are observed exactly. 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 $T^* = \min\{T_1, T_2\}$ 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.

The main methodological contribution of the research presented in this section is to show how parametric assumptions can help tighten the bounds on the object of interest in unidentified competing risks models. This is interesting because the nonparametric bounds that make no assumptions can be quite wide. Since different assumptions will lead to different sets of identified regions, we will consider a number of examples. 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.

4.1 The effect of explanatory variables with parametric restrictions.

We first consider the case where a binary explanatory variable, 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, and the multiplicative effect, α, is the main object of interest. In the next section we also consider the case when no assumption is made on the effect of X on T_2. This model is an example of an accelerated failure time, which is commonly used to describe mortality. It is also 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, let $\varepsilon_k = \log(S_k)$ and consider $\log(T_1)$

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

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

$$X \cdot (\log(\beta) - \log(\alpha)) + (\varepsilon_2 - \varepsilon_1) < 0$$

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 for the explanatory variable and a discrete outcome variable 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).

Because the parameters in (1) are not point–identified, we will construct bounds on them. To do this, 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 particular values of α and β are 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 α and β). The points of support will be denoted by (s_1, s_2), and the associated probabilities by $p(s_1, s_2)$. In this case, the relevant probabilities are

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

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

(corresponding to $X = 0$) and

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

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

(corresponding to $X = 1$).

In order to construct the relevant points of support, 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 1 depict the events in equations (2) and (3), and in equations (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 $\left\{0, \beta^{-1}, 2\beta^{-1}, 3\beta^{-1}, ..., t_{Max}\beta^{-1}\right\}$.

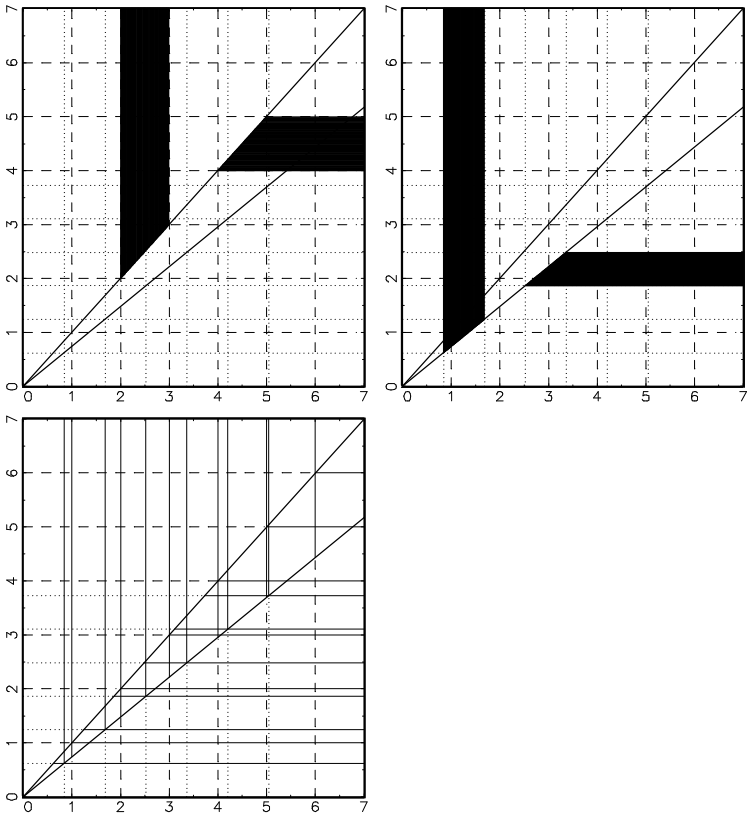

Figure 1: Illustration of Points of Support for Case (a)

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

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

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.

5 Extensions

5.1 No assumption is made on the effect of X on T_2.

It is 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

$$\begin{aligned}
\sum_{\substack{t_k<s_1<t_k+1 \\ s_2>s_1}} p(s_1, s_2) &= P(T = t_k, I = 1 | X = 0), \\
\sum_{\substack{t_k<s_2<t_k+1 \\ s_1>s_2}} p(s_1, s_2) &= P(T = t, I = 0 | X = 0) \\
\sum_{\substack{t_k<as_1<t_k+1 \\ s_2>as_1}} \widetilde{p}(s_1, s_2) &= P(T = t_k, I = 1 | X = 1), \\
\sum_{\substack{t_k<s_2<t_k+1 \\ as_1>s_2}} \widetilde{p}(s_1, s_2) &= P(T = t, I = 0 | X = 1) \\
\sum_{s_1,s_2} p(s_1, s_2) &= 1, \\
\sum_{s_1,s_2} \widetilde{p}(s_1, s_2) &= 1, \\
\sum_{s_2} p(s_1, s_2) &= \sum_{s_2} \widetilde{p}(s_1, s_2) \\
p(s_1, s_2) &\geq 0, \qquad \widetilde{p}(s_1, s_2) \geq 0
\end{aligned}$$

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 4.1, 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.

5.2 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 4.1 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 the points $t_1, t_2, \ldots$ etc.

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 1 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

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

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

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

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

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

subject to the same constraints.

5.3 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 the same way that the Peterson bounds were constructed, 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 2.[6] Most of the points of support are self–explanatory. There are, however, two main differences relative to the points of support in the first panel of Figure 1. The first is that for each region that includes $T_1 = t$ in its interior, one must allow for one point to the left of t and one to the right. The second complication is that for each region, one must allow for a point of support corresponding to each of the discrete values of T_2 that fall in the region. This is needed because one needs these to enforce the restriction that the marginal distributions of T_2 are the same in the two periods. Except for that, the points of support are as they would be in the first panel of Figure 1.

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

$$\min_{\{v_i\},\{p(\cdot,\cdot)\},\{\widetilde{p}(\cdot,\cdot)\}} \sum_{s_1 \leq t} \widetilde{p}(s_1,s_2) - \sum_{s_1 \leq t} p(s_1,s_2)$$

[6] Figure 2 is drawn for the case where the observations are censored after 9 periods and $t = 5.5$.

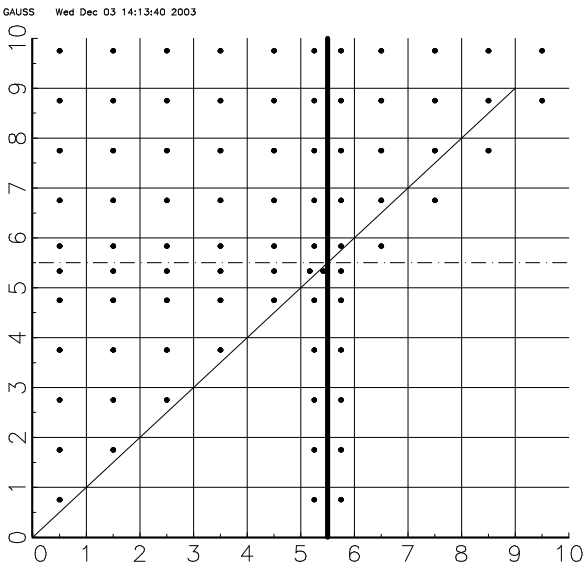

Figure 2: Illustration of Points of Support Necessary to Deal with Exclusion Restrictions

subject to

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

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

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

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

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

$$\max_{\{v_i\},\{p(\cdot,\cdot)\},\{\widetilde{p}(\cdot,\cdot)\}} \sum_{s_1\leq t} \widetilde{p}\left(s_1,s_2\right)-\sum_{s_1\leq t} p\left(s_1,s_2\right)$$

subject to the same constraints.

As in section 5.2, there is no guarantee that the sample analogs of these will be consistent estimators of the lower and upper bounds for $P\left(T_1\leq t\,|X=1\right)-P\left(T_1\leq t\,|X=0\right)$ as the sample analogs of the constraints may have no solution. To derive consistent estimators of these, first define $\widehat{f}$ by

$$\widehat{f}=\max_{\{v_i\},\{p(\cdot,\cdot)\},\{\widetilde{p}(\cdot,\cdot)\}} \sum -v_i$$

subject to

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

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

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

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

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

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

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

$$\min_{\{v_i\},\{p(\cdot,\cdot)\},\{\widetilde{p}(\cdot,\cdot)\}} \sum_{s_1 \leq t} \widetilde{p}(s_1,s_2) - \sum_{s_1 \leq t} p(s_1,s_2)$$

subject to

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

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

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

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

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

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

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

5.4 Bounds with Continuous Covariates or Non–grouped 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). It is therefore not know whether the parameters of the resulting models are point–identified. In this section we demonstrate that it is in principle easy to derive expressions for the identified region for those parameters (whether or not this is a single point).

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

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

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

$$\begin{aligned}
\int_{t_k/a(X)}^{t_{k+1}/a(X)} \int_{a(X)s_1/b(X)}^{\infty} p(s_1, s_2)\, ds_2 ds_1 &= P(T = t_k, I = 1|\, X), && (14)\\
\int_{t_k/b(X)}^{t_{k+1}/b(X)} \int_{b(X)s_2/a(X)}^{\infty} p(s_1, s_2)\, ds_1 ds_2 &= P(T = t_k, I = 0|\, X\ ), && (15)\\
\int\int p(s_1, s_2)\, ds_1 ds_2 &= 1, && (16)\\
p(s_1, s_2) &\geq 0 && (17)
\end{aligned}$$

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

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

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.

6 The Change between 1970 and 2000 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.

6.1 Results assuming independence

As a baseline, we start by constructing bounds for the time dummy under the assumption that the time dummy has a (different) multiplicative effect on the duration until death for both cancer and CVD and imposing independence. In order to provide a fair comparison with the rest of our results, we use the identical estimation method, except that we estimate the bounds for the parameters separately rather than jointly. The conclusions from this estimation should be qualitatively similar to the conclusions we reach by looking at the raw hazard rates: the main differences here are that improvements are expressed in terms of increases in the time until death rather than in decreases in the hazard rates, that we impose a multiplicative functional form and that we treat the data as grouped.

We compute bounds for four demographic groups separately, and for three different periods, 1970 to 1980, 1970 to 1990, and 1970 to 2000. Recall that 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 will signal improvements. The results are in Table 2. Not surprisingly the results show large improvements from 1970 to 2000 in CVD for for all groups: the duration until death from CVD increased between 30 to 40% relative to 1970. On the other hand, we find a very small, albeit positive, improvement in cancer for all groups. The survival until death increased by about 6% for white men during the same period, by about 9% for white women, and it increased by about 2% for black men and women.

For completeness, and for future reference, Table 2 also includes the results for cancer exclusive of lung cancer and from lung cancer alone. These are given in the last two rows for each panel.

6.2 Main Results

We now present our main results which construct bounds without assuming independence, as in section 4.1. We do assume that the potential duration to death from other causes is independent of the potential duration until death from cancer and the potential duration until death from CVD.

The results are in Table 3. For all groups we find that the CVD duration increased substantially

from 1970 to 2000, 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 by 2000, certainly smaller than the percentage increases for CVD, but not negligible. However for all 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 every decade.

We compare these results with those we presented in the previous section (Table 2). The coefficients for CVD are similar with or without independence, especially for white men, but the estimated improvements are larger when we do not assume independence. On the other hand the coefficients for cancer are much larger when we do not assume independence: the improvements more than double for all groups.

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. This is true whether or not we assume independence.

6.3 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 public policies is to decrease mortality, this is the metric we want to use to calculate the benefits of policies to fight cancer. We estimate these counterfactuals as described in Section 5.2. Since we have censored the data at age 80 (and the model is likely to be unreliable in the tail of the distribution), we consider the effect on the probability of surviving past age 75.

The results are presented in Table 4. In the first row for each group we report the fitted probability of surviving past age 75 in 1970. In the second row we report probability of surviving past age 75 in the absence of any progress in cancer (but including progress in CVD) and in the third row we report this probability in the absence of progress in CVD (but including progress in cancer). Finally in the fourth row we report the fitted probability of surviving past 75 in 2000.

The difference between the fourth and the first row represents the total increase in the probability of survival past 75 from all causes from 1970 to 2000. In the case of white males, this probability increased by about 24 percentage points, from 40.8% to 64.6%. From row 2 we see

that in the absence of cancer progress this probability would have been 62.3% in 2000 rather than 64.6%, a difference of 2.3 percentage points. Therefore from this vantage point progress in cancer accounts for about 10% of the total increase in survival.

Alternatively we can look at what the probability of survival would have been in the absence of CVD progress by looking at the third row. In the absence of CVD progress survival rates would have been about 43.1% rather than 40.8%, therefore we also find that for white males cancer progress accounts for about for 10% of the total increase in survival in this period. Similar calculations for other groups show that cancer progress accounts for about 15% of the total increase in survival for white women, 8% for black women and 6.7% for black men. The results are almost identical irrespective of the counterfactual we use.

It is clear that one cannot estimate the counterfactual effect on the life expectancy without additional assumptions about the tail behavior of the distribution of the latent durations. On the other, one might be interested quantities such as the percentiles of the duration until death. Bounds on such quantities can be calculated from bounds one the counterfactual survivor functions. These are depicted in Figure 7.

Next we present results where we impose large progress in CVD to predict what the contributions in mortality from cancer will be as CVD progress continues into the future. The estimates above give us a sense of what the contribution of progress in cancer has been up to 2000. But to the extent that these benefits will continue into the future these estimates of the benefits are too low. Using the estimated probabilities (which need not be unique for a given point of support), we can ask what the contribution of cancer will be once the time until death from CVD has increased by for example, 1.7 (relative to its 1970 value) which is roughly an additional 30% increase (compared to 2000). Row 4 of Table 4 present the probability of surviving past age 75 in the absence of progress in cancer but given this large progress in CVD. This probability is about 71.7% for white males in 2000. In the last row we report the probability that includes progress in cancer at the rate we estimate from 1970 to 2000. It is 74%, or 2.3 percentage point higher than in the absence of cancer progress and 9.4 percentage points higher that the probability given progress in both today. Therefore the progress in cancer we have made so far will account for about 25% of the additional increase in survival past age 75 (2.3/9.4=25) when survival from CVD reaches 1.7 of its 1970 value. Similar calculations suggest that cancer progress will account for about 28% of the additional increases in survival past age 75 for white women, 9.5% for black males and 15% for black women.

7 Estimation issues

7.1 Specification checks

Because 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, we may be interested in estimating separate cancer trends, for cancer excluding lung and for lung cancer only.[7] As mentioned, the results assuming independence for cancer excluding lung are in Table 2. The estimated cancer trends are somewhat larger if we exclude lung cancer (around 7-9% for men), especially for women (compared to Table 2). In Table 5 we present bounds for all cancers excluding lung cancer without assuming independence. We find much larger improvements in cancer when we exclude lung for all groups. The trends are about twice as large as those that include lung, about 19 and 46% for white men and women respectively, and 9 and 45% for black men and women. Again these improvements are much larger than we estimated in Table 2 when we assumed independence.

Because lung cancer and CVD have a common risk, smoking,[8] it may be incorrect to include lung cancer with the third cause of death which we treat as independent. So we re-estimate non-lung cancer trends by grouping all other causes of death into the "other" category, including CVD. (Notice that this is not necessarily correct either since it estimates a single trend for all other causes of death.) Our results (Table 6) 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.

Since smoking affects both CVD and lung cancer, relaxing the independence assumption should result in drastically different results for lung cancer. In the last row of each panel of Table 2, we report the results from estimating trends in lung cancer assuming independence. These results suggest that there was a significant decline in the time until death since 1970 for all groups and in all periods. The results without independence are in Table 7. Once we account for the (known) dependence between CVD and lung cancer, we find that there has been significant progress in lung cancer for all groups except black females (compared to Table 2). 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

[7] 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)).

[8] See references in next section.

this period for all groups, by as much as 15%. Only for black women do we still find a decrease in the duration once we account for dependence, but again this decrease is smaller than if we assume independence. These results suggest trends in lung cancer estimated under independence are substantially biased.

Generally speaking these specification checks do suggest that it may not be appropriate to estimate trends for cancer as a whole, but rather that it would be preferable to separate cancers. Conceptually it is straightforward to extend our method to estimate trends for more than two causes of death without assuming independence. It would also be straightforward to include additional categorical covariates. However both of these extensions are computationally difficult, so we have not pursued them here.

Interestingly, excluding or including lung cancer has only a small effect in our estimates of CVD progress. The imposition of independence does not greatly affect the trends either, even though our results do suggest that cancer and CVD are dependent. Intuitively this occurs 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 smaller ones.

Another important limitation of our estimation method 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 WWW 5.1). These results are presented in Table 9. 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.

7.1.1 Some Data Issues

There are several data issues in calculating age-specific mortality rates using matched data from the census and the death certificate files that are potentially problematic because they may affect our trend estimates.

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. 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. Furthermore, the extent of the undercount varies with the census year (Schenker (1993)). This problem is again larger for blacks than for whites.

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)). 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. Overall age misreporting appears to be a very important issue mostly among blacks. These data issue may explain why some of our results are not very robust for blacks and suggest that our results for blacks must be taken with caution.

Another issue is whether causes of death are correctly specified in the death certificate.[9] More importantly the issue is whether there have been significant changes from 1970 to 2000 in the 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)). Overall changes in the observed causes of death have not significantly changed overtime for broad causes of death.

[9]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.

7.1.2 Additional evidence

Our findings provide support for the claim that there has been progress in cancer, measured in terms of the increases in the underlying cause-specific duration. In this section we provide evidence from other sources consistent with our findings.

We looked for any evidence that there were indeed innovations in terms of cancer 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[10]), 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 showed 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.[11] Breast cancer treatment also changed in the 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, Feuer, Harlan, Wun, Johnson, and Abrams (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)).

In spite of the fact that this evidence is consistent with our trend estimates, it is worth keeping in mind that the trends that we estimate can also reflect changes in lifestyle and demographic characteristics, some of which may reflect prevention, and some which may be completely unrelated to scientific advances in cancer. Ultimately we cannot say with certainty that the trends we estimate

[10]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.

[11]A review of the evidence by the U.S. Preventive Services Task Force is available at http://www.ahrq.gov/clinic/3rduspstf/breastcancer/brcanrr.htm#ref4

are uniquely related to progress in treatment or whether they also reflect prevention and cohort effects.

8 Conclusions

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 make no parametric assumptions on the nature of the dependence between risks, and consider an accelerated failure time model with categorical covariates and grouped durations. Because this model is not point identified, we estimate bounds for the effects of the categorical covariates.

We use our method to estimate changes in cancer and cardiovascular mortality since 1970. The estimated bounds for the effect of time on the duration until death for either cause are extremely tight, much tighter than the bounds one can obtain without making any assumptions at all (Peterson (1976)). Such bounds can therefore be obtained under many more situations and making fewer assumptions than the previous literature has suggested.

Previous research has estimated trends in cancer mortality by assuming independence and has found little or no progress. We find that trends in cancer show much larger improvements than previously estimated. We find that time until death from cancer increased by about 10% for white males and 20% for white women from 1970 to 2000 for all cancers, and by about 19% for white males and 40% for white women if we exclude lung cancer. These estimates are more than twice as large as estimates derived under independence. These improvements are not all due to changes in smoking for younger cohorts. Also we find that not all improvements took place in the 1990s; for women, we find significant improvements going back to the 1970s. Our counterfactuals suggest that cancer improvements have accounted for about 10-15% of the increase in the probability of surviving past age 75 since 1970 among whites. Although less robust, we find similar results for blacks.

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.

Chiang, C. L. (1991): "Competing Risks in Mortality Analysis," *Annual Reviews Public Health*, 12, 281–307.

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., 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).

Llorca, J., and M. Delgado-Rodriguez (2001): "Competing Risks Analysis using Markov Chains: Impact of, Cerebrovascular and Ischaemic Heart Disease in Cancer Mortality," *International Journal of Epidemiology*, 30, 99–101.

Lynch, H. T., 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.

Mariotto, A., E. J. Feuer, L. C. Harlan, L.-M. Wun, K. A. Johnson, and J. Abrams (2002): "Trends in Use of Adjuvant Multi–Agent Chemotherapy and Tamoxifen for Breast Cancer in the United States 1975-1999," *Journal of the National Cancer Institute*, 94(21).

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.

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.

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.

Rothenberg, R. B. (1994): "Competing Mortality and Progress against Cancer," *Epidemiology*, 5, 197–203.

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).

SEER (2004): "Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 9 Regs Public-Use, Nov 2003 Sub (1973-2001)," *National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2004.*

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.

Wohlfart, J., and P. K. Andersen (2001): "Commentary: Secular Trends in the Context of Competing, Risks," *International Journal of Epidemiology*, 30, 102–103.

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

9 Appendix

9.1 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

Death rate–causes of death classification

Deaths from cardiovascular diseases 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".

9.2 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 $\{\theta_1^{\min} - 0.05, \theta_1^{\min} - 0.04, \theta_1^{\min} - 0.03, ..., \theta_1^{\max} + 0.08\} \times \{\theta_2^{\min} - 0.05, \theta_2^{\min} - 0.04, \theta_2^{\min} - 0.03, ..., \theta_2^{\max} + 0.08\}$.

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 third grid is then given by $\{\theta_1^{\min} - 0.01, \theta_1^{\min} - 0.009, \theta_1^{\min} - 0.008, ..., \theta_1^{\max} + 0.015\} \times$
$\{\theta_2^{\min} - 0.01, \theta_2^{\min} - 0.009, \theta_2^{\min} - 0.008, ..., \theta_2^{\max} + 0.015\}$.

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 Assuming Independence

Results for White Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.126, 1.129)	(1.239, 1.250)	(1.392, 1.400)
Coefficient on Cancer	(1.001, 1.029)	(1.001, 1.029)	(1.059, 1.060)
Coef. on Cancer (excl lung)	(1.091, 1.093)	(1.126, 1.129)	(1.075, 1.076)
Coef. on Lung Cancer	(0.910, 0.911)	(0.905, 0.909)	(0.968, 0.968)

Results for White Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.091, 1.093)	(1.201, 1.206)	(1.286, 1.291)
Coefficient on Cancer	(1.001, 1.029)	(1.059, 1.060)	(1.087, 1.093)
Coef. on Cancer (excl lung)	(1.091, 1.093)	(1.236, 1.250)	(1.334, 1.346)
Coef. on Lung Cancer	(0.843, 0.852)	(0.849, 0.852)	(0.840, 0.851)

Results for Black Males

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.126, 1.129)	(1.201, 1.206)	(1.316, 1.320)
Coefficient on Cancer	(0.972, 0.999)	(0.965, 0.965)	(1.001, 1.029)
Coef. on Cancer (excl lung)	(1.084, 1.090)	(1.091, 1.093)	(1.091, 1.093)
Coef. on Lung Cancer	(0.847, 0.848)	(0.847, 0.851)	(0.847, 0.852)

Results for Black Females

	1970–1980	1970–1990	1970–2000
Coefficient on CVD	(1.160, 1.166)	(1.273, 1.280)	(1.334, 1.346)
Coefficient on Cancer	(1.001, 1.029)	(0.972, 0.999)	(1.001, 1.029)
Coef. on Cancer (excl lung)	(1.059, 1.060)	(1.126, 1.129)	(1.239, 1.250)
Coef. on Lung Cancer	(0.840, 0.846)	(0.840, 0.842)	(0.851, 0.852)

TABLE 3: Marginal Identified Regions without Assuming Independence

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 4: Counterfactual Probability of Surviving Age 75

Results for White Males

	1970–1980	1970–1990	1970–2000
No Progress	(0.402 , 0.402)	(0.407 , 0.407)	(0.408 , 0.408)
Progress in CVD	(0.488 , 0.488)	(0.575 , 0.575)	(0.623 , 0.623)
Progress in Cancer	(0.402 , 0.402)	(0.407 , 0.407)	(0.431 , 0.431)
Progress in Both	(0.488 , 0.488)	(0.579 , 0.579)	(0.646 , 0.646)
$a = 1.7$, $b = 1$	(0.733 , 0.733)	(0.711 , 0.711)	(0.717 , 0.717)
$a = 1.7$, $b = estimate$	(0.733 , 0.733)	(0.716 , 0.716)	(0.740 , 0.740)

Results for White Females

	1970–1980	1970–1990	1970–2000
No Progress	(0.586 , 0.586)	(0.588 , 0.588)	(0.589 , 0.589)
Progress in CVD	(0.655 , 0.655)	(0.703 , 0.703)	(0.734 , 0.734)
Progress in Cancer	(0.597 , 0.597)	(0.610 , 0.610)	(0.616 , 0.616)
Progress in Both	(0.666 , 0.666)	(0.725 , 0.725)	(0.760 , 0.760)
$a = 1.7$, $b = 1$	(0.822 , 0.822)	(0.823 , 0.823)	(0.828 , 0.828)
$a = 1.7$, $b = estimate$	(0.833 , 0.833)	(0.845 , 0.845)	(0.855 , 0.855)

TABLE 4 (cont.): Counterfactual Probability of Surviving Past 75

Results for Black Males

	1970–1980	1970–1990	1970–2000
No Progress	(0.337 , 0.337)	(0.345 , 0.345)	(0.348 , 0.348)
Progress in CVD	(0.399 , 0.399)	(0.442 , 0.442)	(0.500 , 0.500)
Progress in Cancer	(0.342 , 0.342)	(0.355 , 0.355)	(0.359 , 0.359)
Progress in Both	(0.404 , 0.404)	(0.452 , 0.452)	(0.511 , 0.511)
$a = 1.7$, $b = 1$	(0.627 , 0.627)	(0.616 , 0.616)	(0.615 , 0.615)
$a = 1.7$, $b = estimate$	(0.632 , 0.632)	(0.626 , 0.626)	(0.626 , 0.626)

Results for Black Females

	1970–1980	1970–1990	1970–2000
No Progress	(0.465 , 0.465)	(0.467 , 0.467)	(0.472 , 0.472)
Progress in CVD	(0.548 , 0.548)	(0.594 , 0.594)	(0.638 , 0.638)
Progress in Cancer	(0.474 , 0.474)	(0.486 , 0.486)	(0.487 , 0.487)
Progress in Both	(0.557 , 0.557)	(0.613 , 0.613)	(0.653 , 0.653)
$a = 1.7$, $b = 1$	(0.737 , 0.737)	(0.740 , 0.740)	(0.736 , 0.736)
$a = 1.7$, $b = estimate$	(0.746 , 0.746)	(0.759 , 0.759)	(0.751 , 0.751)

TABLE 5: Marginal Identified Regions Excluding Lung Cancer

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 6: Marginal Identified Regions
Excluding Lung Cancer and Aggregating All Other Courses of Death

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 7: Marginal Identified Regions for Lung Cancer

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	$(0.879, 0.886)$	$(1.160, 1.166)$	$(0.945, 0.959)$

TABLE 9: 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	$(0.556, 2.284)$	$(0.644, 2.230)$	$(0.702, 2.332)$

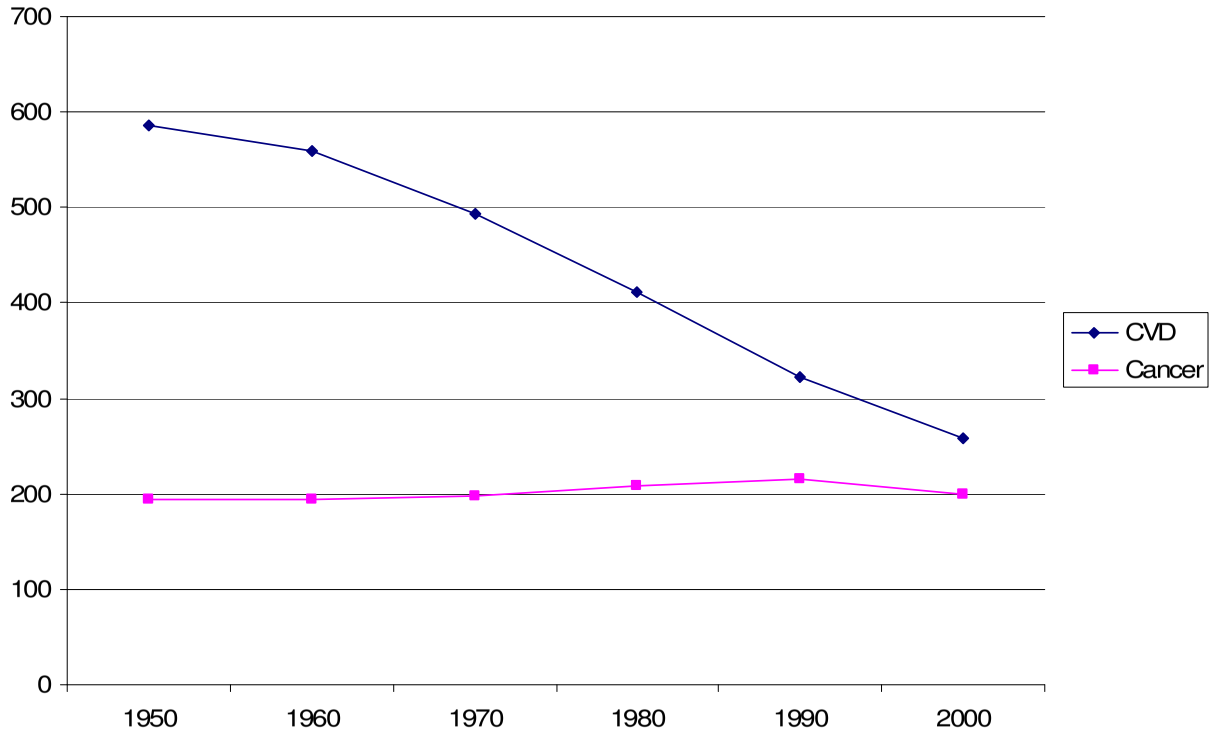

Figure 3: Trends in age–adjusted mortality 1950–2000 (all persons)

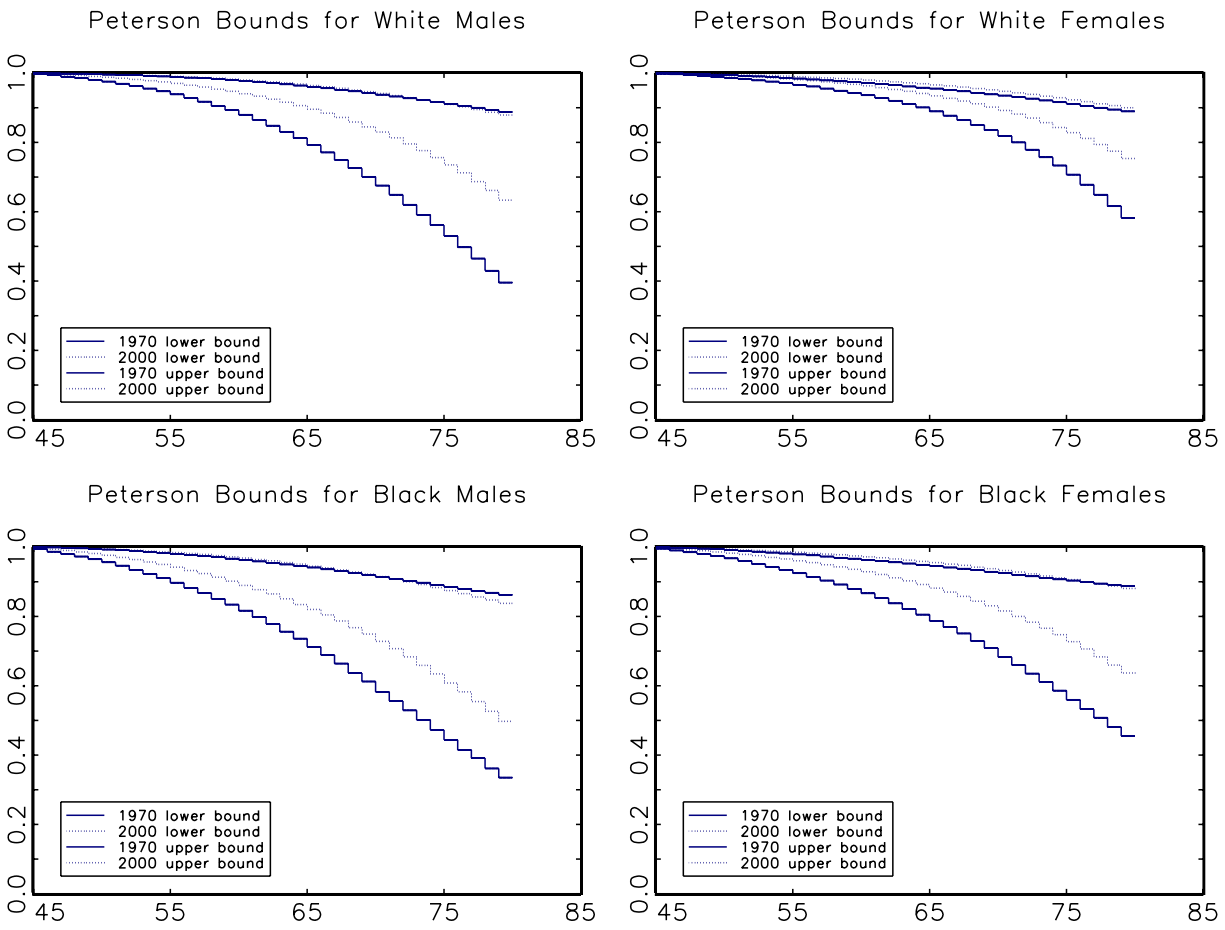

Figure 4: Peterson Bounds on the Survivor Functions in 1970 and 2000

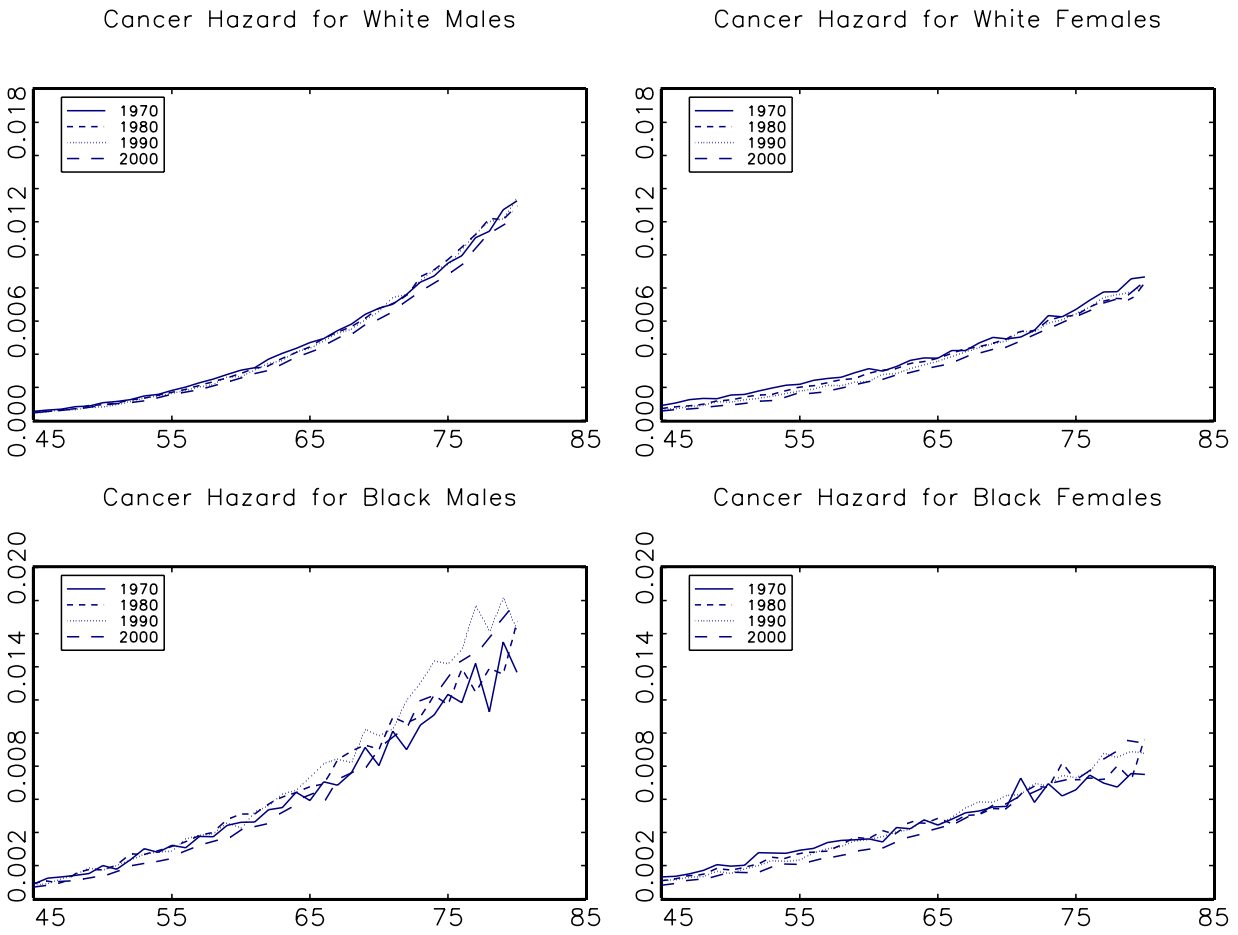

Figure 5: Hazard Rates for the Cancer

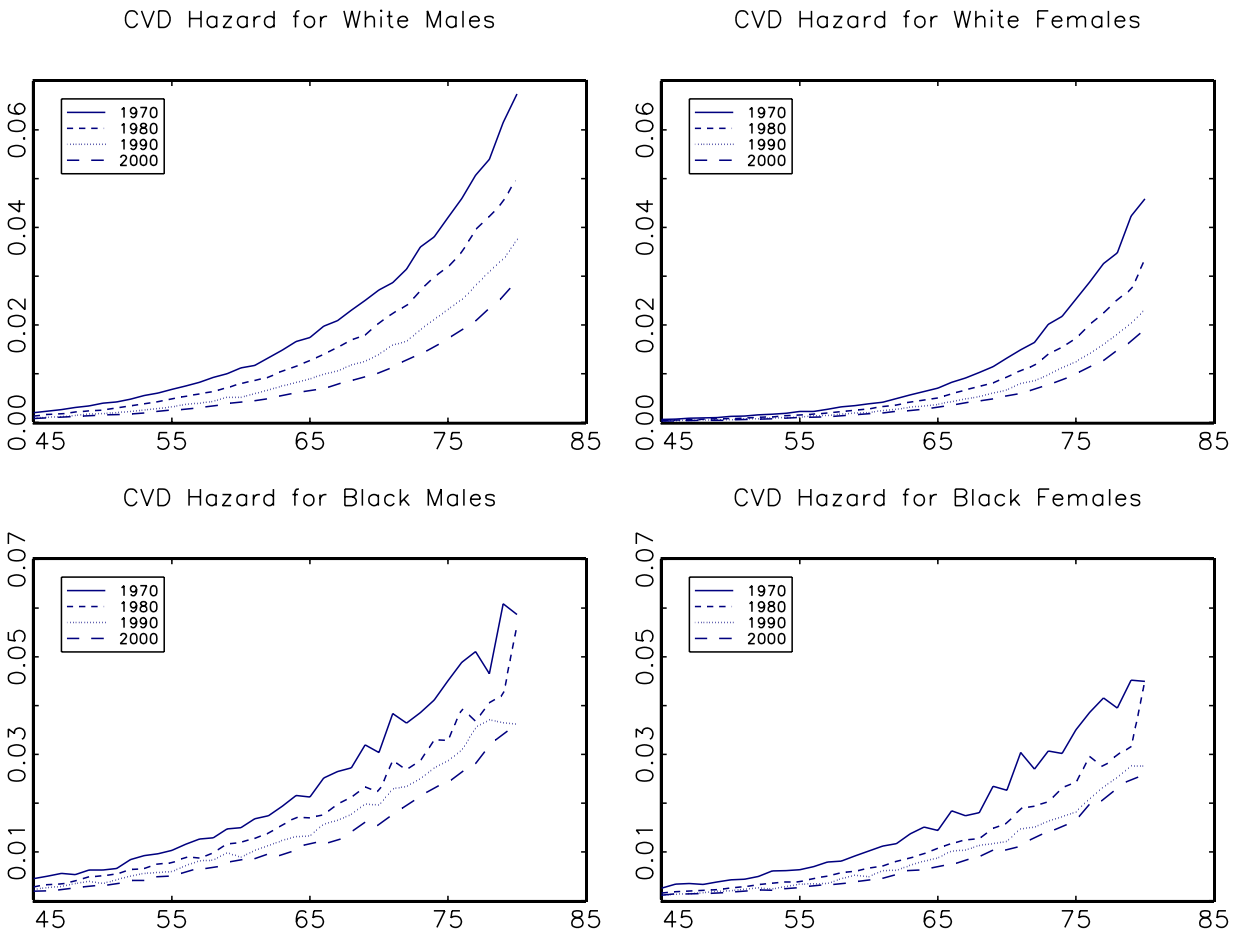

Figure 6: Hazard Rates for the CVD

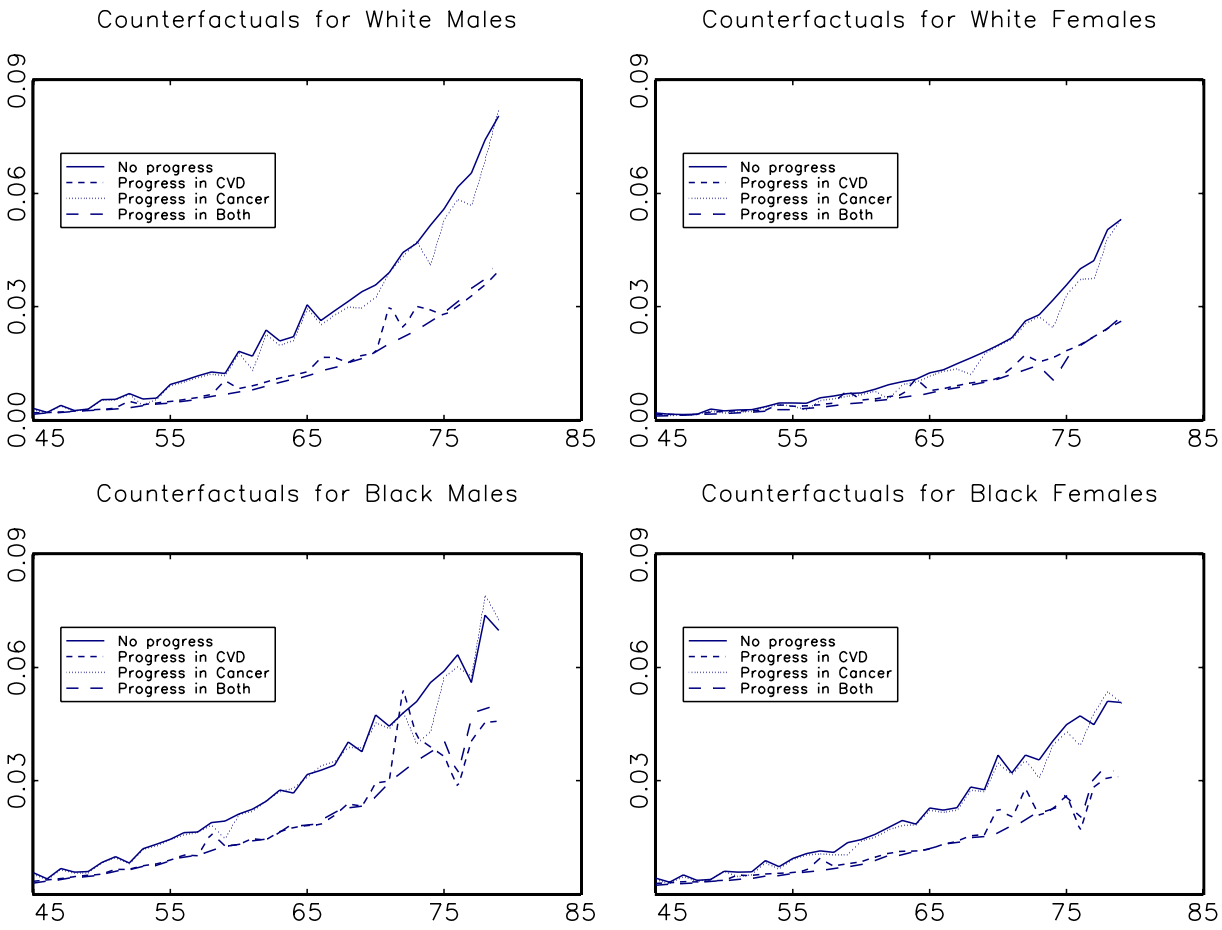

Figure 7: Counterfactual Policy Evaluations