Bounds in Competing Risks Models and the War on Cancer*

Bo E. Honoré† Adriana Lleras-Muney‡

This Version: October 2005

Abstract

In 1971 President Nixon declared war on cancer. Thirty years later, many have declared the war a failure: the age-adjusted mortality rate from cancer in 2000 was essentially the same as in the early 1970s. Meanwhile the age-adjusted mortality rate from cardiovascular disease fell dramatically. Since the causes underlying cancer and cardiovascular disease are likely to be correlated, the decline in mortality rates from cardiovascular disease may in part explain the lack of progress in cancer mortality. Because competing risks models (used to model mortality from multiple causes) are fundamentally unidentified, it is difficult to get a clear picture of the trends in cancer. 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 weak parametric assumptions in order to obtain identification. We provide a framework to estimate competing risk models with interval outcome data and discrete explanatory variables, both of which are common in empirical applications. 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 fairly tight and suggest much larger improvements in cancer than previously estimated.

Keywords: Bounds, Competing Risks, Cancer, Cardiovascular.

JEL Classification: I10, C40.

---

*We would like to thank Jaap Abbring, Josh Angrist, Eric J. Feuer, Marco Manacorda, researchers at the National Cancer Institute, and seminar participants at CAM at the University of Copenhagen, London School of Economics, Massachusetts Institute of Technology, NBER Summer Institute, Princeton University, University College London, 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 (Lleras–Muney) and by the 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 (Honoré).

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

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

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 mortality from cancer increased from 198.7 (per 100,000) in 1973 to 213 in 1993, and 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, as shown in Figure 1.[2] It has been hypothesized that the decline in mortality rates from cardiovascular disease is somewhat responsible for the lack of progress 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.[3] 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

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

[2] These are trends in age-adjusted mortality, and therefore they include mortality rates from all ages. Age-adjusted mortality is the weighted average of age-specific mortality rates, where the weights are given by the 2000 age-specific population to avoid comparisons that reflect changes in the age distribution over time. These trends are well-known and have been documented elsewhere, e.g. Howe, Wingo, Thun, Ries, Rosenberg, Feigal, and Edwards (2001).

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

assumptions. Most importantly, we do not assume that the underlying risks are independent, but instead impose relatively weak parametric assumptions in order to obtain identification. The theoretical contribution of the paper is to provide a framework for estimating competing risk models with interval data and discrete explanatory variables, both of which are common in empirical applications.[4] There are a number of 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).

Using mortality data from the US grouped by gender, race, single year of age, and by cause of death in 1970, 1980, 1990 and 2000, we use the proposed method to estimate the group–specific trends in cancer mortality.[5] We find that trends in cancer show much larger improvements than previously estimated using similar data and the same covariates. The main empirical contribution of this paper is to show that allowing for dependence across risks can result in very different estimates of the trends in cancer.

We view these trends as reflecting "progress against cancer" broadly defined, including advances in prevention, diagnosis, and treatment. We cannot and do not attempt to separate these effects– the purpose of looking at mortality trends as a measure of progress is in fact to capture all of these effects (see footnote 5). However, given the data available, it is worth noting that we cannot

[4] A number of other papers have used a similar approach to construct bound on objects of interest in semi– or nonparametric models. See for example Peterson (1976), Manski (1990), Balke and Pearl (1997), Horowitz and Manski (2000), Manski and Tamer (2002), Mahajan (2003) and Molinari (2005).

[5] 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 rates 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 against cancer.

separate our trend estimates from trends in unobserved characteristics (other than gender and race), in particular from other cohort or period effects. Another limitation of this paper is that we rely on the commonly used accelerated failure time model to improve identification. Although our results suggest this model is a good fit for our data, ultimately we cannot test this assumption.

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 we analyze the results separately for each group.

With our data we can calculate the observed hazard rates using a discrete time Kaplan–Meier estimator. Figures 2 and 3 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 (T) 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 (δ). Mathematically, we observe T and δ where

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

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

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

and the objects 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 is that at the level of an individual, the durations are dependent. The fact that there are several known risk factors that are common to both CVD and cancer suggests this may be the case. 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, among others. Excessive alcohol use 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.

Alternatively, heterogeneity across individuals can cause the population 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)).[6] This will cause the latent duration until death from CVD and cancer to be correlated. Furthermore there are 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 who have never married; and Hispanics were less likely than non-Hispanics to drink (Schoenborn, Adams, Barnes, Vickerie, and Schiller (2004)).

Overall this evidence would suggest that the same factors that increase the likelihood of CVD also increase the risk of cancer. Evidence from country level data is also consistent with positive dependence across these two risks. For example, Vaupel and Yashin (1999) report that cancer mortality rates are lowest in countries with the highest CVD rate, suggesting that progress against CVD increases cancer mortality rates.[7]

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

[7] Also it is interesting that, as in United States, mortality rates in Western Europe from cancer increased until the late 1980s and started declining thereafter; while mortality rates from CVD fell since 1970. For trends in cancer in the Europe see Levi, Negri, and La Vecchia (2000) and Levi, Negri, and La Vecchia (2002), and for trends in CVD

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 durations (T_1, T_2), there exists a unique pair of univariate distributions, such that if S_1 and S_2 are two independent random variables with those marginal distributions, 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 conditional 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 on the covariates. The approach taken in this paper is a combination of the first and the third approaches.

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 2 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.[8] 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.

Alternatively, one can make no assumptions on the joint distribution of the underlying durations,

see Levi, Negri, and La Vecchia (2002).

[8]Slud and Byar (1988) provide an example where the effect of a given covariate is constructed to lower the survival rate for a given cause, but the assumption of independence makes it appear as if the same factor actually increases the survival rates. 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).

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 by age 60, 15% of individuals have died of CVD and 10% have died of cancer. The survival rate[9] 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 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 increased or decreased in this 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 the results may be entirely driven 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 is not necessary 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, 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. However, the durations are observed in groups in many data sets. This raises the question of what can be

[9] The epidemiology literature on cancer often uses the term "survival rate" to refer to the fraction of people who are alive five years after being diagnosed. In this paper we do not condition on diagnosis and we do not only consider five year periods.

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.

It is important to understand that neither we nor Heckman and Honoré (1989) nor Abbring and van den Berg (2003) "solve" the nonidentification of competing risks models. In each of the cases, identification is studied under a set of additional assumptions. Without such assumptions, it is not possible to improve upon the Peterson bounds. In our application, we assume a specific functional form for the marginal distributions, and it is that functional form that allows us to tighten the bounds on the parameters of interest. Exclusion restriction assumptions would be an appealing alternative, but we were unable to find plausible restrictions for the specific application in this paper. Instead we make use of the accelerated failure time model to improve identification. This is one of the most commonly used models in epidemiology, the main alternative being the proportional hazard model. To our knowledge, neither of these has been given a solid theoretical foundation in the medical literature.[10] Rather, their popularity seems to stem from the fact that they deliver good fits for describing all-cause mortality rates.[11] Our main motivation for considering the accelerated failure time model is its ease of interpretation and the close connection between it and linear regression models with sample selection.

As mentioned above, 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 of 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

[10] This is unlike the situation in economics where theory sometimes motivates restrictions that can be used to tighten the nonparametric bounds. See for example Blundell, Gosling, Ichimura, and Meghir (2004).

[11] Depending on the population, time period, and cause of death that studies used, they have found one or the other specification to be a better alternative. Many authors note that the accelerated failure time model model has a more intuitive and natural interpretation than the proportional hazards model, but both models implicitly make parametric assumptions which may not appropriately describe the underlying unknown distributions.

$t_k < T^* \leq t_{k+1}$ 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,

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

where (S_1, S_2) is independent of X, and the multiplicative effects, α and β, are the main objects of interest. In the next section we also consider the case where no assumption is made on the effect of X on T_2. This model is an example of an accelerated failure time model, which is commonly used to describe mortality. It was originally introduced by Cox (1972), who gave it a physical interpretation in the context of mortality. From the equivalence between proportional hazard models with Weibull baseline hazards and Weibull accelerated failure time models, it follows that a model where the marginals obey a Weibull proportional hazard assumption is consistent with our functional form assumption. 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 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 see that one can obtain bounds on the parameters under (1), consider a simple case in which observations are censored after 2 time–periods, and one observes

$$\begin{aligned} P(T=0, I=0|\, X=0) &= P(T=1, I=0|\, X=0) = \\ P(T=0, I=1|\, X=0) &= P(T=1, I=1|\, X=0) = \\ P(T=2|\, X=0) &= \frac{1}{5} \end{aligned}$$

If $\alpha = \beta = 2$ then this has a number of implications for the distribution of (T, I) when $X = 1$. For example, all observations that were censored when $X = 0$ will still be censored when $X = 1$, as will observations with $T = 1$ (so T^* is between 1 and 2). On the other hand, none of the observations with $T = 0$ (so T^* is between 0 and 1) will be censored. This means that the probability of censoring must be 0.6 when $X = 1$. So, if one observes that the probability of censoring when $X = 1$ is 0.5, then one would rule out $\alpha = \beta = 2$. Of course, in this case there are additional constraints, for example $P(T = 0 \,\text{or}\, 1, I = 0|\, X = 1) = 0.2$. The main insight in this section is to keep track of all such implications for a given (α, β). 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

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

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

(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 S_1 is concerned**,** in the sense that two values of S_1 that fall in the same interval will lead to the same observed value of T_1. 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 S_2.

The first two panels in Figure 5 depict the events in equations (2) and (3), and in equations (4) and (5), respectively. The dashed lines in the panels correspond 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 $\left\{0, \beta^{-1}, 2\beta^{-1}, 3\beta^{-1}, ..., t_{Max}\beta^{-1}\right\}$. Panel 1 of Figure 5 depicts the situation when $X = 0$, whereas Panel 2 depicts the situation when $X = 1$. The vertical shaded areas in both panels correspond to the case when individuals are observed to die from the first cause of death (depicted in the x–axis) in a particular interval. Since any point in this area will result in the same observed outcome (when either $X = 0$ or $X = 1$), there is no loss of generality in assuming that (S_1, S_2) has a discrete distribution over that area with arbitrarily located points of support. Likewise, the horizontal shaded areas correspond to the case when individuals are observed to die from the second cause of death (depicted in the y–axis) in a particular interval, and again there is no loss of generality in assuming that (S_1, S_2) has a discrete distribution over that area with arbitrarily located points of support.

Panel 3 of Figure 5 combines the first two panels. Each polygon (depicted in solid lines) contains only values of (S_1, S_2) that will lead to the same observed values of T_1 and T_2, both when $X = 0$ and $X = 1$. 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 these polygons.

The identified region for (α,β) is the set of (a,b) such that there exists $p(s_1,s_2)$ satisfying[12]

$$\begin{aligned}
\sum_{\substack{t_k<s_1<t_{k+1}\\ s_2>s_1}} p(s_1,s_2) &= P(T=t_k, I=1|\, X=0), && (6)\\
\sum_{\substack{t_k<s_2<t_{k+1}\\ s_1>s_2}} p(s_1,s_2) &= P(T=t_k, I=0|\, X=0\ ), && (7)\\
\sum_{\substack{t_k<as_1<t_{k+1}\\ bs_2>as_1}} p(s_1,s_2) &= P(T=t_k, I=1|\, X=1\ ), && (8)\\
\sum_{\substack{t_k<bs_2<t_{k+1}\\ as_1>bs_2}} p(s_1,s_2) &= P(T=t_k, I=0|\, X=1\ ), && (9)\\
\sum_{s_1,s_2} p(s_1,s_2) &= 1, \qquad p(s_1,s_2)\geq 0 && (10)
\end{aligned}$$

(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 (2004) and Molinari (2005), 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 suggested in Honoré and Tamer (2004), 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

$$f(a,b) = \max_{\{v_i\},\{p(\cdot,\cdot)\}} \sum -v_i \tag{11}$$

subject to

$$\begin{aligned}
v_k + \sum_{\substack{t_k<s_1<t_{k+1}\\ s_2>s_1}} p(s_1,s_2) &= P(T=t_k, I=1|\, X=0) \qquad k=1,\ldots M, && (12)\\
v_{M+k} + \sum_{\substack{t_k<s_2<t_{k+1}\\ s_1>s_2}} p(s_1,s_2) &= P(T=t_k, I=0|\, X=0\ ) \qquad k=1,\ldots M, && (13)
\end{aligned}$$

[12] Imposing $\beta=1$ in this example, will give the identified region for α, under the exclusion restriction that X has no effect on T_2.

$$v_{2M+k}+\sum_{\substack{t_k<as_1<t_{k+1}\\ bs_2>as_1}} p(s_1,s_2) = P(T=t_k, I=1|\, X=1\ ) \qquad k=1,\ldots M, \tag{14}$$

$$v_{3M+k}+\sum_{\substack{t_k<bs_2<t_{k+1}\\ as_1>bs_2}} p(s_1,s_2) = P(T=t_k, I=0|\, X=1\ ) \qquad k=1,\ldots M, \tag{15}$$

$$v_{4M+1}+\sum_{s_1,s_2} p(s_1,s_2) = 1, \qquad p(s_1,s_2)\geq 0 \qquad \text{for all } (s_1,s_2), \tag{16}$$

$$v_i \geq 0 \qquad k=1,\ldots 4M+1 \tag{17}$$

It is clear that this approach generalizes to the case where there are more than two latent failure times, and to the case where the (vector of) explanatory variable(s) takes more than two values.

Theorem 2 (and the accompanying corollary) of the appendix establishes 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. Moreover, the uniform rate of convergence equals that of $\widehat{P}$ to P. It therefore follows by the argument in Manski and Tamer (2002) that the identified region can be consistently estimated by 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 that converges to 0 more slowly than the rate of convergence of $\widehat{P}$.

The consistency argument with the corresponding rate of convergence is quite generic. For the particular problem studied in this paper, it is possible to establish additional results based on the following Lemma, which is proved in the Appendix.

Lemma 1 *The functions $f(a,b)$ and $\widehat{f}(a,b)$ are both piecewise constant over the same finite number of regions.*

Lemma 1 essentially makes the parameter space discrete with a finite number of elements.

Note that the setup in (11) forces one to underestimate all the probabilities. While this does not affect the consistency of the resulting estimator of α and β, it may be intuitively unappealing. It might therefore be more attractive to consider an alternative linear programming problem that allows the errors to be positive or negative. The disadvantage of this approach is that it increases the dimensionality of the linear programming problem, and therefore we do not use it for our main results. This alternative linear programming problem is fully specified in the Appendix (section 9.6) and is implemented only as a sensitivity check.

5 Extensions

5.1 More General Functional Forms

It is easy to generalize the setup in section 4.1 to cover functional forms that are not linear. Specifically, consider

$$(T^*, I) = \begin{cases} (\min\{S_1, S_2\}, 1\{S_1 < S_2\}) & \text{for } X = 0, \\ \\ (\min\{g_1(S_1;\alpha), g_2(S_2;\beta)\}, 1\{g_1(S_1;\alpha) < g_2(S_2;\beta)\}) & \text{for } X = 1, \end{cases} \tag{18}$$

where g_1 and g_2 are known functions and α and β are unknown parameters. In this case the identified region for (α, β) is the set of (a, b) that maximize the linear programming problem

$$f(a, b) = \max_{\{v_i\},\{p(\cdot,\cdot)\}} \sum -v_i \tag{19}$$

subject to (12)–(17) but with as_1 and bs_2 replaced by $g_1(s_1; a)$ and $g_2(s_2; b)$, respectively. We pursue a version of this in our empirical application.

5.2 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

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

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 $p\left(s_1, s_2\right)$ and $\widetilde{p}\left(s_1, s_2\right)$ satisfying

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

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 $\left(S_1, S_2\right)$ 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. We also implement this in our application.

5.3 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 if only the distribution of T_1 had changed.

Consider for example the setup on section 4.1 and define

$$\widetilde{T}_1^* = \min\left\{\alpha S_1, S_2\right\}$$

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 $X = 1$ in order to find the effect that X has on T through T_2 alone (keeping the distribution of T_1 where it would be when $X = 1$).[13] It might also be interesting to know the effect of eliminating a risk altogether. This suggests comparing the distribution of T^* given $X = 1$ to the distribution of

$$\widetilde{T}_2^* = \alpha S_1$$

(or the distribution of T^* given $X = 0$ to the distribution of S_1).

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 $X = 0$ or given $X = 1$. It is therefore natural to also consider the distribution of the grouped version of $\widetilde{T}_1^*$ or T_2^*. 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 $p(\cdot, \cdot)$ we have

$$\begin{aligned} P\left(\widetilde{T}_1^* < t_k\right) &= P\left(\min\left\{\alpha S_1, S_2\right\} < t_k\right) \\ &= \sum_{s_1 < t_k/\alpha \text{ or } s_2 < t_k} p(s_1, s_2) \end{aligned}$$

The last expressions are affected by the fact that the points of support are not uniquely determined. Before proceeding, it is therefore necessary to consider every polygon depicted in solid lines in the third graph of Figure 5 and determine whether the location of a point within the region changes whether the event $\{s_1 < t_k/\alpha \text{ or } s_2 < t_k\}$ occurs. If it does, then one must place two points of support in the region corresponding to whether or not $\{s_1 < t_k/\alpha \text{ or } s_2 < t_k\}$.

One can then calculate population bounds on $P\left(\widetilde{T}_1^* < t_k\right)$ by minimizing and maximizing (over a and b) the function $\sum_{s_1 < t_k/\alpha, s_2 < t_k} p(s_1, s_2)$ subject to (6)–(10). Unfortunately, the sample analog of this (which replaces $P(T = t_k, I = i|\, X = x)$ by $\widehat{P}(T = t_k, I = i|\, X = x)$) will not produce a consistent estimator of the upper and lower bounds on $P\left(\widetilde{T}^* < t_k\right)$, because there is no guarantee that the sample version of (6)–(10) 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 (11). The reason for this is that for a 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

$$f(a, b) = \max_{\{v_i\}, \{p(\cdot,\cdot)\}} \sum -v_i \tag{20}$$

[13] Other effects of this type could be considered. For example, one could compare the distribution of

$$\min\{S_1, \beta S_2\}$$

to the distribution of T^* given $X = 0$ in order to find the effect that X has on T through T_2 alone (keeping the distribution of T_1 where it would be when $X = 0$).

subject to the constraints (12)–(17) except that $P(\cdot)$ has been replaced with $\widehat{P}(\cdot)$. $\widehat{\Theta}$ is then an estimate of the identified region for (α, β). 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 (12)–(17) (except that $P(\cdot)$ has been replaced with $\widehat{P}(\cdot)$) and subject to the constraint that

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

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.

The same setup can be used to construct estimates of the upper and lower bounds for objects related to life expectancy. Specifically consider the censored duration until death, $\min\{T, M\}$, rounded down to the nearest integer. The upper bound for this is obtained by maximizing $g(a, b)$ over (a, b) in $\widehat{\Theta}$ where

$$g(a, b) = \max_{\{v_i\}, \{p(\cdot,\cdot)\}} \sum \operatorname{int}\{\min a s_1, s_2, M\} \cdot p(s_1, s_2)$$

subject to the same constraints. The lower bound is obtained by minimizing the same function subject to the same constraints.

5.4 Other extensions

We considered additional extensions, which we only briefly summarize here since they are not used in the empirical application below. A more detailed discussion of how to implement these can be found in Honoré and Lleras-Muney (2004).

The first extension considers how bounds could be constructed if exclusion restrictions were available. 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. 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. If such restrictions could be used, then one could relax the multiplicative assumption in the competing risk model. This is essentially done in the same way that the Peterson bounds were constructed, but with the added restriction that the marginal distribution for one of the durations is the same in the two subsamples given by $X = 0$ and $X = 1$.

Another interesting extension is to consider estimation when the explanatory variable of interest is continuous or the durations are observed exactly. Here, we have focused on the case where the explanatory variable X is discrete and the durations are grouped, because this corresponds to the situation on our empirical application. This is the case in which the competing risks 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.

6 The Change between 1970 and 2000 in 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 construct bounds under the commonly used assumption of independence. In doing so, we assume that the time dummy has a (different) multiplicative effect on the duration until death for both cancer and CVD. 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. Except for the fact that we account for the grouping, this amounts to estimating a standard accelerated failure time model for each survival time. For details on the estimation, see the appendix. 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 all groups: the duration until death from CVD increased between 30% and 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 from cancer increased by about 6% for white men during the same period, by about 9% for white women, and 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.[14] 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% to 20% for women by 2000, certainly smaller than the percentage increases for CVD, but not negligible. However for white males the increase was mostly concentrated in the 1990s; from 1970 to 1990 the increases were small, about 2% to 4%. The same is not true for females, who saw some significant improvements in every decade. For black males there were very small improvements in each 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. The fact that the bounds presented in Table 3 do not contain the bounds in Table 2 as subsets suggests that the multiplicative model with independence is

[14]The theory presented earlier requires one to define the interval estimates as the set of parameter values for which the function value is within some ε_n of its maximum, where ε_n is a sequence of numbers that converges to 0 more slowly than $\sup_{a,b}\left|\widehat{f}(a,b) - f(a,b)\right|$. Since measurement error is likely to be more important than estimation uncertainty, we ignore this issue in the results presented in this paper. Consider, for example the improvement for white males between 1970 and 2000. This is based on approximate 40,000,000 observations in each of the two years. A small simulation study that draws observations using the sample probabilities, suggests that the average sup–difference is approximately 0.00057 with a standard deviation of approximately 0.00016 (the maximum value over 1000 replcations was 0.001). If we choose ε_n to be the mean plus five standard deviations (0.00217), then the interval estimates for a and b would change to $(1.34, 1.39)$ and $(1.11, 1.25)$, respectively. While these intervals are somewhat larger than the intervals presented in Table 3, the substantive conclusions based on them are not different.

inconsistent with the observed data.[15]

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 either in the absence of improvements for cardiovascular disease, or with the observed improvements in CVD. We estimate these counterfactuals as described in Section 5.3. 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, and on life expectancy (censored at 80). These numbers can be used to evaluate the progress that had been obtained thus far. Second we ask what the changes in mortality would be if we could eliminate cancer as a cause of death. This will give an upper bounds on the gain from further progress in the fight on cancer.

The results are presented in Table 4. In the first row for each group we report the actual probabilities in each period. In the next row we report (bounds for) the fitted probability of surviving past age 75 in 1970. These fitted values are based on estimation using data from 1970 and 1980, 1970 and 1990, and 1970 and 2000, respectively. This explains why one should not expect the numbers in this row to be identical, although we would have considered it as evidence against the multiplicative function form if they had differed by a great deal. Similarly in the fifth row we report the fitted probability of surviving past 75 in 1980, 1990 and 2000, respectively. Comparing the fitted values in rows two and five to the actual values in the first row reveals that the fitted values are consistently very close to the actual values. This provides some evidence that our functional form assumption is not inconsistent with the data. It is also worth noting that the fitted values are always consistently below the actual values. This should be expected since our linear programming setup always underestimates the probabilities (see the end of section 4).

In the third row of Table 4 we report (bounds for) the probability of surviving past age 75 in the absence of any progress in cancer (but including progress in CVD), and in the next row, we report this probability in the absence of progress in CVD (but including progress in cancer).

In the case of white males, the probability of surviving past age 75 increased by about 19.5

[15] This is confirmed by the fact that the optimal function values (not reported here) for the optimization problem in section 9.5 are quite far from 0.

percentage points, from 56.1% in 1970 to 75.6% in 2000. From row 3 we see that, in the absence of cancer progress, this probability would have been between 66% and 73.8% in 2000. Therefore from this vantage point progress in cancer ranges from 2 to 10.6 percentage points and accounts for somewhere between 10% and 55% 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 fourth row. In the absence of CVD progress survival rates would have been between 57.4% and 59.5% rather than 57.4%; therefore we find that for white males cancer progress accounts for about for 0-11% of the total increase in survival in this period.

The difference in the estimates is driven by the choice of baseline: the first estimate uses the later year as a baseline and therefore computes counterfactuals that allow progress in CVD. The second set of estimates uses 1970 as a baseline and is computed in the absence of progress in CVD. The difference tells us about the extent to which progress in cancer (in the form of reductions in cancer mortality) cannot occur without concurrent progress in CVD (and vice-versa), which depends on the extent to which the diseases are correlated (at the extreme, if the diseases were very highly correlated, eliminating one would have almost no noticeable impact on mortality because everyone would die seconds later).

Similar calculations for other groups show that cancer progress accounts for 22%-100% of the total increase in survival for white women using 2000 as a base, or 0% to 25% using 1970; for black women the range is 3.7% to 40% in 2000 and 0% to 11% in 1970. Finally for black men cancer progress accounts for 5%-59.6% of improvements in 2000, but only for 0 % to 11.8% in 1970.

It is clear that the bounds on the changes discussed above are potentially too wide as they are based on a comparison of bounds for two parameters. Alternatively we can estimate bounds on the change directly. These alternative bound estimates are in Table 5. The bounds that are estimated directly are somewhat tighter, although the general pattern and the qualitative conclusions do not change.

Table 4 also presents estimates of the effect of eliminating either cancer or CVD on the probability of surviving past age 75. These can be used to estimate upper bounds on the potential benefits of additional medical innovation in one of the two causes of death, without any changes in the alternative cause of death. One conclusion to be drawn from these results is that, for males in 1970, reductions in CVD would have much larger impacts in the overall survival compared to cancer. In 2000 improvements in cancer and CVD have comparable potential benefits. For women, the results are somewhat different in that the potential gains from improvements in cancer and CVD are comparable throughout the period.

We also use our model to calculate changes in life expectancy conditional on survival to age 45 (and with censoring at 80). The results presented in Table 6 are very similar to those in Table 4, except that we can express changes in the survival distribution in terms of additional years of life, which can be used in cost benefit analysis if we have estimates of the dollar value of an additional year of life. For example, for white males the actual increase from 1970 to 2000 is approximately 3.3 years. In the absence of progresses in cancer, progress in CVD disease would account for 0.9 to 2.6 years. One estimate of the progress in cancer is therefore the remaining 0.7 to 2.4 years of life. Given the progress that occurred in CVD, the maximum increase that could have occurred is between 3.3 and 4.3. For white females, the comparable bound is between 1.6 and 2.8 years, so the potential gains from improvements in cancer appear smaller than those for white males. This is somewhat unintuitive since cancer is a relatively larger risk for females than for males. However this is partially due to the fact that the parameter of interest is lifetime censored at 80. Since whites females live the longest (see table 1) the censoring results in lower progress estimates for women. For this reason we prefer the estimates in Table 4.

It is worth noting that the fitted values in this table are always lower than the actual values. Again, this can be explained by the fact that we underestimate all the probabilities, and they therefore do not add up to one.

7 Estimation issues

7.1 Functional Form Assumption

The strongest identifying assumption that we make is that the trends for a given disease are the same for all individuals ages 45 to 80. We have relaxed this assumption by estimating models that allow for different trends for younger and older individuals. For the second period, the latent duration is specified as follows:

$$T = \begin{cases} a_1 S & \text{if } S \leq t_0, \\ a_1 t_0 + a_2(S - t_0) & \text{if } S > t_0. \end{cases}$$

We estimate this model for $t_0 = 15$ (corresponding to a break at age 60, since we start at age 45) or $t_0 = 20$ (break at age 65). These break points split the data into two groups of approximately equal size, and they are also commonly used in epidemiology.[16] In this specification the effects are no longer multiplicative for the older age group. The estimates for a_1 and a_2 are presented in Tables 7A and 7B. The overall qualitative conclusions from this, are similar to the conclusions we

[16]For example, the Centers for Disease Controls and Prevention reports mortality trends by cause every year for all, and by age grouping individuals ages 45 to 64, and those ages 65 and above (CDC 2002, 2003, 2004). See also Doll (1991) or Chiang (1991).

reached when pooling all ages. For all demographic groups, and regardless of which of the break points we choose, there is evidence of significant progress in cancer by 2000. For males (white or black), this progress appears to be concentrated in the 1990s,[17] whereas for women (white or black) progress in cancer occurred every decade.

The function value, reported for each group, suggests that these models that allow for breaks fit the data better (the function value is closer to zero), and furthermore, that the break at 60 is a better fit than the break at 65. However what is not clear is how much of a difference it makes to allow for these breaks in terms of the overall amount of progress for the entire group during a given period: none of these coefficients are directly comparable to those in Table 3, which pool all ages together; neither are the results in both tables directly comparable to each other. In order to gauge the extent to which this alternative specification affects the overall conclusions, we graph the predicted duration in 2000 as a function of the survival in 1970 by cause and for each group, using our estimates. Figures 6 and 7 show the results. It is clear from the figures that for all models and groups there is substantial progress in both CVD and cancer between 1970 and 2000. Moreover it makes very little difference whether we allow for breaks or not. Overall our qualitative conclusions are unaffected when we use this alternative specification. For this reason, and for ease of interpretation, we estimate the original model in the rest of the paper.

The results presented so far suggest that the model is a good fit for the data. The results are not particularly sensitive to adding breaks. Below, we show they are robust to using a different starting age. In our counterfactual exercises we found that the fitted values predicted by the model are remarkably close to the actual values. Also the maximized function value is close to zero (although without inference theory, these findings are only suggestive). Thus our choice of functional form appears to be reasonable, although we re–iterate that competing risks models are fundamentally unidentified, so other functional forms could also be consistent with the observed data.

7.2 Other 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 trends for all cancers except lung cancer.[18] As mentioned earlier, the results assuming

[17] Although there appears to have been some progress in the 1970s for white males in the model that uses 65 as a break.

[18] Deaths from lung cancer diminished in the 1990s because of decreases in smoking that started to take place in the 1960s and that are most likely unrelated to progress in prevention and treatment since 1973 (Andersen, Remington,

independence for cancer excluding lung cancer are in Table 2. Assuming independence, the estimated trends are somewhat larger if we exclude lung cancer (around 7-9% for men), especially for women. In Table 8, we present the same bounds without assuming independence. We find much larger improvements when we exclude lung cancer for all groups. The trends are about twice as large as those that include lung cancer, 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 those in Table 2 (when we assumed independence). 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 could re-estimate non-lung cancer trends by grouping all other causes of death into the "other" category, including CVD. But notice that this would not be correct either since it estimates a single trend for all other causes of death.

This suggests that it may not be appropriate to estimate trends for cancer as a whole, but rather that it would be preferable to separate cancers. As mentioned earlier, it is conceptually 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 as both the number of constraints and number of unknown parameters in the linear programming problem increase linearly in the number of causes of death and in the number of different values of the covariates. We have therefore not pursued them here (except to the extent that we estimate separate models based on gender and race).

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 5.2). These results are presented

Trentham-Dietz, and Reeves (2002)).

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 difficult 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.

If progress in detection and treatment prevents some deaths prior to age 45 in 2000 but not in 1970, then our choice to restrict the sample to those ages 45 and above could bias our results. There are two reasons why we present results only conditional on survival until age 45. The first is that cancers that affect children and young adults are considered to be very different from those that affect mature adults. The trends for cancer would therefore be expected to be quite different for these age groups. The second and main reason for the age restriction is that the death rates below age 45 were low for both causes of death in all years (see Figures 8 and 9). However the graphs suggest that mortality rates started rising between ages 40 and 45. Thus we re-estimated the results in Table 3 starting with ages 40 (instead of 45). The results are presented in Table 10. As would be expected, the coefficients are somewhat different, but the general patterns are very similar. Overall, we find substantial progress in cancer from 1970 to 2000, and the timing of this progress is as described in the main result section.

Finally, we did some of the calculations using the alternative formulation of the linear programming problem given in (22) (section 9.6) which does not systematically underestimate the probabilities. As discussed, this is intuitively more appealing as it does not systematically underestimate the probabilities. The results from this are very close to those obtained from those reported here. For example, the estimated intervals for the coefficients for the change between 1970 and 2000 for white males changed from $(1.389, 1.391)$ to $(1.392, 1.400)$ for CVD and from $(1.134, 1.153)$ to $(1.134, 1.142)$ for cancer.

7.3 Some Data Issues

There are several potential problems in calculating age-specific mortality rates using matched data from the census and the death certificate files that 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 age of 65, who tend to understate their age. 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 issues 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.[19] 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 previous research suggests that changes in the observed causes of death have not significantly changed over time for broad causes of death.

7.4 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

[19]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.

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[20]), namely 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 over 50.[21] Breast cancer treatment also changed in the 1980s with the dissemination of adjuvant chemotherapy, including multi-agent chemotherapy and tamoxifen. Additional changes in treatment 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)).

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

[20]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.

[21]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

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. 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.

Balke, A., and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92(439), 1171–1176.

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.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2004): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," IZA Discussion Paper NO. 1350.

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.

——— (1972): "Regression models and life-tables," *J. Royal Statistical Soc. Series B-statistical Methodology*, 34.

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 A. Lleras-Muney (2004): "Bounds in Competing Risks and the War on Cancer," National Bureau of Economic Research WP 10963.

Honoré, B. E., and E. T. Tamer (2004): "Bounds on Parameters in Dynamic Discrete Choice Models," Princeton University.

Horowitz, J. L., and C. F. Manski (2000): "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95(449), 77–??

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).

Levi, F., L., F. E. Negri, and C. La Vecchia (2000): "The Decline in Cancer Mortality in the European Union," *European Journal of Cancer*, 36(15), 1965–1968.

——— (2002): "Trends in Mortality from Cardiovascular and Cerebrovascular Diseases in Europe and Other Areas of the World," *Heart*, 88, 119–124.

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.

Mahajan, A. (2003): "Misclassified Regressors in Binary Choice Models," Stanford University.

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).

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

Molinari, F. (2005): "Partial Identification of Probability Distributions with Misclassified Data," Cornell University.

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 abdand 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 Estimates 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 Heterogenous 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 Uniform convergence of objective function for linear programming

Theorem 2 *Suppose that a function $f(\theta; P)$ is defined by*

$$f(\theta; P) = \max_{\{p_m\},\{v_i\}} \sum_i -v_i \tag{21}$$

subject to

$$\begin{aligned} P(j) - \sum_{m=1}^{M} p_m \pi_m(j;\theta) &= v_j \qquad \text{for all} \qquad j \in \mathcal{J} \\ 1 - \sum_{m=1}^{M} p_m &= v_0 \\ p_m &\geq 0 \qquad \text{for all} \qquad 1 \leq m \leq M \\ v_j &\geq 0 \qquad \text{for all} \qquad j \in \mathcal{J} \end{aligned}$$

where $P()$ is a vector of probabilities and the functions $\pi_m(j;\theta)$ non–negative and bounded from above. Then

$$|f(\theta; P_1) - f(\theta; P_2)| \leq const \cdot (1 - c_1) + const \cdot (1 - c_2)$$

where $c_1 = \min_j \frac{P_2(j)}{P_1(j)}$, $c_2 = \min_j \frac{P_1(j)}{P_2(j)}$ and the constants are independent of P_1, P_2 and θ.

Proof: Consider a particular value of θ.

For that θ, the linear problem clearly has a feasible solution for any vector of probabilities, P (namely $v_j = P(j)$ for $j \in \mathcal{J}$, $v_0 = 1$ and $p_m = 0$ or $m = 1, ..., M$). Now consider a vector of probabilities, P_1, and let (p^1, ν^1) be the maximizer for (21) with $P = P_1$.

We will now construct a feasible solution to (21) with $P = P_2$.

Define $p_m^* \equiv c_1 p_m^1 \geq 0$ where $c_1 = \min_j \frac{P_2(j)}{P_1(j)} \leq 1$. Then $p_m^1 - p_m^* = (1 - c_1) p_m^1$.

Also define

$$\begin{aligned}
v_j^* &\equiv P_2(j) - \sum_{m=1}^{M} p_m^* \pi_m(j;\theta) \qquad \text{for all } j \in \mathcal{J} \\
&= P_2(j) - c_1 \left(\sum_{m=1}^{M} p_m^1 \pi_m(j;\theta) \right) \\
&\geq \frac{P_2(j)}{P_1(j)} P_1(j) - \frac{P_2(j)}{P_1(j)} \left(\sum_{m=1}^{M} p_m^1 \pi_m(j;\theta) \right) \\
&= \frac{P_2(j)}{P_1(j)} v_j \geq 0
\end{aligned}$$

$$v_0^* \equiv 1 - \sum_{m=1}^{M} p_m^* = 1 - c_1 \sum_{m=1}^{M} p_m^1 \geq 1 - \sum_{m=1}^{M} p_m^1 \geq 0$$

$(\{p^*\}, \{\nu^*\})$ is a feasible solution to (21) with $P = P_2$. Moreover

$$\left| v_j^* - v_j \right| \leq |P_2(j) - P_1(j)| + M(1 - c_1) \leq (1 + M)(1 - c_1)$$

(since all the π's are between 0 and 1) and

$$|v_0^* - v_0| \leq M(1 - c_1)$$

so $f(\theta; P_2) \geq f(\theta; P_1) - const \cdot (1 - c_1)$. Interchanging P_1 and P_2 we have $f(\theta; P_1) \geq f(\theta; P_2) - const \cdot (1 - c_2)$ with $c_2 = \min_j \frac{P_1(j)}{P_2(j)}$

We therefore have that

$$const \cdot (1 - c_2) \geq f(\theta; P_2) - f(\theta; P_1) \geq -const \cdot (1 - c_1)$$

and $|f(\theta; P_2) - f(\theta; P_1)| \leq const \cdot (1 - c_1) + const \cdot (1 - c_2)$.

Corollary 3 *Consider a vector of probabilities P such that $\min_j P(j) > 0$. Then*

$$\sup_{\theta} \left| f\left(\theta; \widehat{P}\right) - f(\theta; P) \right| = O_p\left(\widehat{P} - P\right).$$

9.2 Proof of Lemma 1.

The proof is broken into two steps.

First consider a particular choice of (a, b), (a_1, b_1). See Figure 5. The linear programming problem (11) assigns probabilities to each region of the figure for (a_1, b_1). Now imagine that one perturbs (a, b) to (a_2, b_2). Unless new regions are created by this or existing regions disappear, location of the regions will move continuously with (a, b). In other words, one can (uniquely) associate the regions created by (a_2, b_2) with the regions created by the original values (a_1, b_1) (provided that no new regions were created and no existing regions eliminated). One can therefore consider the feasible solution to the linear programming problem for (a_2, b_2) that leaves the probabilities for each region at the values that were optimal for (a_1, b_1). This will leave the objective function at (a_2, b_2) at the optimal level for (a_1, b_1). It therefore follows that $f(a_1, b_1) \leq f(a_2, b_2)$. By interchanging (a_1, b_1) and (a_2, b_2), it follows that $f(a_1, b_1) = f(a_2, b_2)$ for two sets of parameter values (a_1, b_1) and (a_2, b_2) as long as the change from (a_1, b_1) to (a_2, b_2) moves the regions without eliminating or creating any. Since the regions in Figure 5 are constructed without reference to the probabilities associated with them, it follows that $\widehat{f}(\cdot, \cdot)$ is also the same for two sets of parameter values (a_1, b_1) and (a_2, b_2) as long as the change from (a_1, b_1) to (a_2, b_2) moves the regions without eliminating or creating any.

In the second step of the proof, we will now argue that as one varies (a, b) regions will be created or eliminated only a finite number of times. This will establish that f and $\widehat{f}$ are piecewise constant over the same finite number of regions.

Consider the figure. The solid lines are the ones that are independent of (a, b). Step one of the proof implies that the function value only changes when the configuration of (a, b) moves the new lines in a way that creates or eliminates region. It is convenient to reparameterize from (a, b) to $\left(\frac{a}{b}, b\right)$. As one varies $\frac{a}{b}$ from 0 to ∞, the line $s_2 = \frac{a}{b} s_1$ will touch the intersection of a vertical and horizontal solid line only a finite number of times. Now consider a particular value of $\frac{a}{b}$ and consider the effect of varying b (holding $\frac{a}{b}$ fixed). Each value of b (along with the implied value of a) will allocate each of the points in $\left\{\frac{1}{b}, \frac{2}{b}, ..., \frac{T_{\max}}{b}, \frac{1}{a}, \frac{2}{a}, ..., \frac{T_{\max}}{a}\right\}$ into one of the sets $]0, 1[$, $[1, 1]$, $]1, 2[$, $[2, 2]$, $]2, 3[$, $[3, 3]$,... $]T_{\max} - 1, T_{\max}[$, $[T_{\max}, T_{\max}]$. Since the number of points in $\left\{\frac{1}{b}, \frac{2}{b}, ..., \frac{T_{\max}}{b}, \frac{1}{a}, \frac{2}{a}, ..., \frac{T_{\max}}{a}\right\}$ is finite and the number of sets into which they are allocated are finite, it follows that this particular $\frac{a}{b}$ will generate only a finite number of regions.

9.3 The Data

9.3.1 Population data

We obtained population numbers as of April 1st of each decennial years, by single year of age, gender and race. These data come from April 1st population counts or estimates from the Census Bureau, from the following sources:

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

2. The 1980 population estimates are found at

   http://www.census.gov/population/estimates/nation/e80s/e8081rqi.txt

3. The 1990 population estimates were found at

   http://www.census.gov/population/estimates/nation/e90s/e9090rmp.txt

4. The 2000 population counts are available at the US Census Bureau's website: http://factfinder.census.gov/servlet. The white population counts obtained from Census table PCT12A, Black population counts from table PCT12B and total population counts from PCT12.

9.3.2 Mortality data

The number of deaths comes from the Multiple Cause of Death files for 1970, 1980, 1990 and 2000. 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."

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

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.

9.5 Estimation under independence

To estimate the parameters under independence, we first estimate the marginal distribution of T_1 and T_2 using a Kaplan–Meier estimator. We then estimate bounds for a by

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

subject to

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

where the points of support are determined in a way that is similar to the way we did it without independence. Bounds for b are estimated analogously.

9.6 Alternative Linear programming problem

Note that the setup in (11) forces one to underestimate all the probabilities. While this does not affect the consistency of the resulting estimator of α and β, it may be intuitively unappealing. It might therefore be more attractive to consider the linear programming problem

$$f(a, b) = \max_{\{v_i\},\{u_i\},\{p(\cdot,\cdot)\}} \sum -(v_i + u_i) \tag{22}$$

subject to

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

The disadvantage of this approach is that it increases the dimensionality of the linear programming problem. In our application, we therefore focus on the first formulation.

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	72.00	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	78.80	80.20
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	69.40	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

	1970–1980	1970–1990	1970–2000
White Males			
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	$(0.910, 0.911)$	$(0.905, 0.909)$	$(0.968, 0.968)$
White Females			
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	$(0.843, 0.852)$	$(0.849, 0.852)$	$(0.840, 0.851)$
Black Males			
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	$(0.847, 0.848)$	$(0.847, 0.851)$	$(0.847, 0.852)$
Black Females			
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

	1970–1980	1970–1990	1970–2000
White Males			
Coefficient on CVD	$(1.126, 1.129)$	$(1.295, 1.296)$	$(1.389, 1.391)$
Coefficient on Cancer	$(1.001, 1.029)$	$(1.020, 1.035)$	$(1.134, 1.153)$
Function Value	-0.019	-0.036	-0.037
White Females			
Coefficient on CVD	$(1.091, 1.093)$	$(1.158, 1.160)$	$(1.236, 1.238)$
Coefficient on Cancer	$(1.091, 1.093)$	$(1.154, 1.160)$	$(1.201, 1.206)$
Function Value	-0.009	-0.016	-0.021
Black Males			
Coefficient on CVD	$(1.126, 1.129)$	$(1.201, 1.206)$	$(1.334, 1.346)$
Coefficient on Cancer	$(1.030, 1.034)$	$(1.063, 1.066)$	$(1.072, 1.074)$
Function Value	-0.005	-0.029	-0.039
Black Females			
Coefficient on CVD	$(1.158, 1.160)$	$(1.231, 1.235)$	$(1.334, 1.346)$
Coefficient on Cancer	$(1.096, 1.096)$	$(1.167, 1.172)$	$(1.160, 1.160)$
Function Value	-0.029	-0.035	-0.050

TABLE 4: Counterfactual Probability of Surviving Age 75

	1970–1980	1970–1990	1970–2000
White Males			
Change in actual prob.	0.561 – 0.636	0.561 – 0.707	0.561 – 0.756
No Progress (fitted in 70)	(0.567, 0.567)	(0.572, 0.573)	(0.574, 0.574)
Progress in CVD	(0.634, 0.643)	(0.699, 0.719)	(0.661, 0.738)
Progress in Cancer	(0.567, 0.572)	(0.572, 0.579)	(0.574, 0.595)
Progress in Both (fitted in end year)	(0.638, 0.643)	(0.707, 0.719)	(0.756, 0.769)
Elim. CVD — Cancer as in end year	(0.638, 0.913)	(0.707, 0.906)	(0.756, 0.919)
Elim. Cancer — CVD as in end year	(0.661, 0.732)	(0.747, 0.812)	(0.785, 0.848)
Elim. CVD — Cancer as in 70	(0.634, 0.909)	(0.699, 0.906)	(0.661, 0.888)
Elim. Cancer — CVD as in 70	(0.567, 0.656)	(0.572, 0.662)	(0.574, 0.663)
White Females			
Change in actual prob.	0.733 – 0.784	0.733 – 0.820	0.733 – 0.843
No Progress (fitted in 70)	(0.736, 0.736)	(0.738, 0.738)	(0.739, 0.740)
Progress in CVD	(0.736, 0.776)	(0.738, 0.802)	(0.739, 0.820)
Progress in Cancer	(0.736, 0.751)	(0.738, 0.760)	(0.739, 0.765)
Progress in Both (fitted in end year)	(0.786, 0.787)	(0.824, 0.826)	(0.843, 0.850)
Elim. CVD — Cancer as in end year	(0.786, 0.917)	(0.824, 0.923)	(0.843, 0.930)
Elim. Cancer — CVD as in end year	(0.786, 0.865)	(0.824, 0.900)	(0.843, 0.916)
Elim. CVD — Cancer as in 70	(0.736, 0.906)	(0.738, 0.900)	(0.739, 0.900)
Elim. Cancer — CVD as in 70	(0.736, 0.824)	(0.738, 0.827)	(0.739, 0.827)
Black Males			
Change in actual prob.	0.473 – 0.540	0.473 – 0.577	0.473 – 0.634
No Progress (fitted in 70)	(0.474, 0.474)	(0.481, 0.482)	(0.486, 0.486)
Progress in CVD	(0.513, 0.542)	(0.536, 0.579)	(0.582, 0.629)
Progress in Cancer	(0.474, 0.481)	(0.481, 0.490)	(0.486, 0.504)
Progress in Both (fitted in end year)	(0.542, 0.542)	(0.587, 0.588)	(0.635, 0.647)
Elim. CVD — Cancer as in end year	(0.542, 0.872)	(0.587, 0.871)	(0.635, 0.880)
Elim. Cancer — CVD as in end year	(0.558, 0.660)	(0.621, 0.708)	(0.678, 0.768)
Elim. CVD — Cancer as in 70	(0.513, 0.872)	(0.536, 0.862)	(0.582, 0.862)
Elim. Cancer — CVD as in 70	(0.474, 0.589)	(0.481, 0.597)	(0.486, 0.599)
Black Females			
Change in actual prob.	0.586 – 0.673	0.586 – 0.713	0.586 – 0.748
No Progress (fitted in 70)	(0.594, 0.594)	(0.598, 0.598)	(0.603, 0.603)
Progress in CVD	(0.616, 0.673)	(0.598, 0.696)	(0.668, 0.740)
Progress in Cancer	(0.594, 0.604)	(0.598, 0.615)	(0.603, 0.622)
Progress in Both (fitted in end year)	(0.679, 0.683)	(0.717, 0.725)	(0.748, 0.764)
Elim. CVD — Cancer as in end year	(0.679, 0.907)	(0.717, 0.909)	(0.748, 0.916)
Elim. Cancer — CVD as in end year	(0.690, 0.774)	(0.726, 0.807)	(0.768, 0.846)
Elim. CVD — Cancer as in 70	(0.616, 0.897)	(0.598, 0.881)	(0.668, 0.891)
Elim. Cancer — CVD as in 70	(0.594, 0.691)	(0.598, 0.694)	(0.603, 0.698)

TABLE 5: Bound on the Change in Counterfactual Probability of Surviving Age 75

	1970–1980	1970–1990	1970–2000
White Males			
Increase in prob. (w. impr in CVD)	$(0.000, 0.027)$	$(0.000, 0.033)$	$(0.027, 0.108)$
Increase in prob. (no impr in CVD)	$(0.000, 0.005)$	$(0.000, 0.011)$	$(0.000, 0.022)$
White Females			
Increase in prob. (w. impr in CVD)	$(0.010, 0.051)$	$(0.021, 0.087)$	$(0.029, 0.111)$
Increase in prob. (no impr in CVD)	$(0.000, 0.015)$	$(0.000, 0.021)$	$(0.000, 0.025)$
Black Males			
Increase in prob. (w. impr in CVD)	$(0.000, 0.029)$	$(0.009, 0.051)$	$(0.016, 0.065)$
Increase in prob. (no impr in CVD)	$(0.000, 0.006)$	$(0.000, 0.008)$	$(0.000, 0.018)$
Black Females			
Increase in prob. (w. impr in CVD)	$(0.009, 0.067)$	$(0.028, 0.127)$	$(0.022, 0.097)$
Increase in prob. (no impr in CVD)	$(0.000, 0.010)$	$(0.000, 0.017)$	$(0.000, 0.019)$

TABLE 6: Counterfactual (Censored)
Life Expectancy at age 45

	1970–1980	1970–1990	1970–2000
White Males			
Change in actual $E[T\|T \geq 45]$.	$72.73 - 74.09$	$72.73 - 75.28$	$72.73 - 76.05$
No Progress (fitted in 70)	$(72.66, 72.71)$	$(72.61, 72.66)$	$(72.63, 72.68)$
Progress in CVD	$(73.46, 73.93)$	$(74.26, 74.95)$	$(73.62, 75.36)$
Progress in Cancer	$(72.66, 72.81)$	$(72.61, 72.78)$	$(72.63, 73.07)$
Progress in Both (fitted in end year)	$(73.89, 73.98)$	$(74.92, 74.98)$	$(75.64, 75.76)$
Elim. CVD — Cancer as in end year	$(73.89, 78.33)$	$(74.92, 78.05)$	$(75.64, 78.22)$
Elim. Cancer — CVD as in end year	$(74.15, 75.42)$	$(75.49, 76.49)$	$(76.06, 77.01)$
Elim. CVD — Cancer as in 70	$(73.46, 78.27)$	$(74.26, 78.02)$	$(73.62, 77.82)$
Elim. Cancer — CVD as in 70	$(72.66, 74.19)$	$(72.61, 74.14)$	$(72.63, 74.16)$
White Females			
Change in actual $E[T\|T \geq 45]$.	$75.67 - 76.51$	$75.67 - 77.09$	$75.67 - 77.46$
No Progress (fitted in 70)	$(75.65, 75.66)$	$(75.63, 75.65)$	$(75.61, 75.63)$
Progress in CVD	$(75.65, 76.27)$	$(75.63, 76.60)$	$(75.68, 76.83)$
Progress in Cancer	$(75.65, 75.91)$	$(75.63, 76.02)$	$(75.61, 76.07)$
Progress in Both (fitted in end year)	$(76.43, 76.47)$	$(76.92, 76.97)$	$(77.25, 77.31)$
Elim. CVD — Cancer as in end year	$(76.43, 78.47)$	$(76.92, 78.46)$	$(77.25, 78.54)$
Elim. Cancer — CVD as in end year	$(76.43, 77.85)$	$(76.92, 78.21)$	$(77.30, 78.44)$
Elim. CVD — Cancer as in 70	$(75.65, 78.27)$	$(75.63, 78.10)$	$(75.68, 78.08)$
Elim. Cancer — CVD as in 70	$(75.65, 77.23)$	$(75.63, 77.21)$	$(75.61, 77.16)$
Black Males			
Change in actual $E[T\|T \geq 45]$.	$70.69 - 71.98$	$70.69 - 72.76$	$70.69 - 73.80$
No Progress (fitted in 70)	$(70.68, 70.69)$	$(70.61, 70.67)$	$(70.62, 70.66)$
Progress in CVD	$(71.39, 71.91)$	$(71.51, 72.42)$	$(72.19, 73.26)$
Progress in Cancer	$(70.68, 70.82)$	$(70.61, 70.90)$	$(70.62, 70.93)$
Progress in Both (fitted in end year)	$(71.93, 71.97)$	$(72.48, 72.63)$	$(73.39, 73.52)$
Elim. CVD — Cancer as in end year	$(71.93, 77.73)$	$(72.48, 77.49)$	$(73.39, 77.54)$
Elim. Cancer — CVD as in end year	$(72.22, 74.00)$	$(72.93, 74.62)$	$(74.08, 75.46)$
Elim. CVD — Cancer as in 70	$(71.39, 77.67)$	$(71.51, 77.27)$	$(72.19, 77.28)$
Elim. Cancer — CVD as in 70	$(70.68, 72.73)$	$(70.61, 72.70)$	$(70.62, 72.65)$
Black Females			
Change in actual $E[T\|T \geq 45]$.	$72.83 - 74.45$	$72.83 - 75.15$	$72.83 - 75.75$
No Progress (fitted in 70)	$(72.71, 72.75)$	$(72.77, 72.80)$	$(72.72, 72.77)$
Progress in CVD	$(73.09, 74.04)$	$(73.01, 74.47)$	$(73.45, 74.96)$
Progress in Cancer	$(72.71, 72.98)$	$(72.77, 73.17)$	$(72.72, 73.13)$
Progress in Both (fitted in end year)	$(74.16, 74.26)$	$(74.80, 74.90)$	$(75.22, 75.35)$
Elim. CVD — Cancer as in end year	$(74.16, 77.96)$	$(74.80, 77.93)$	$(75.22, 77.89)$
Elim. Cancer — CVD as in end year	$(74.29, 75.90)$	$(74.94, 76.41)$	$(75.53, 76.82)$
Elim. CVD — Cancer as in 70	$(73.09, 77.73)$	$(73.01, 77.51)$	$(73.45, 77.51)$
Elim. Cancer — CVD as in 70	$(72.71, 74.56)$	$(72.77, 74.61)$	$(72.72, 74.51)$

TABLE 7A: Marginal Identified Regions (break at age 60)

	1970–1980	1970–1990	1970–2000
White Males			
Coefficient on CVD before age 60	(1.184, 1.199)	(1.378, 1.383)	(1.500, 1.530)
Coefficient on CVD after age 60	(0.950, 0.985)	(0.953, 0.968)	(1.012, 1.091)
Coefficient on Cancer before age 60	(0.953, 0.958)	(1.001, 1.025)	(1.225, 1.240)
Coefficient on Cancer after age 60	(1.160, 1.185)	(1.061, 1.105)	(0.898, 0.977)
Function Value	−0.003460	−0.004795	−0.002246
White Females			
Coefficient on CVD before age 60	(1.169, 1.194)	(1.300, 1.308)	(1.377, 1.381)
Coefficient on CVD after age 60	(0.952, 1.021)	(0.963, 1.023)	(0.952, 1.016)
Coefficient on Cancer before age 60	(1.128, 1.132)	(1.184, 1.194)	(1.255, 1.260)
Coefficient on Cancer after age 60	(0.951, 1.011)	(0.989, 1.058)	(1.011, 1.085)
Function Value	−0.000797	−0.001479	−0.002890
Black Males			
Coefficient on CVD before age 60	(1.147, 1.176)	(1.301, 1.306)	(1.457, 1.457)
Coefficient on CVD after age 60	(1.023, 1.052)	(0.969, 1.028)	(0.940, 1.014)
Coefficient on Cancer before age 60	(1.004, 1.029)	(1.004, 1.019)	(1.126, 1.126)
Coefficient on Cancer after age 60	(0.936, 1.015)	(0.956, 1.015)	(0.944, 1.013)
Function Value	−0.000768	−0.001481	−0.006133
Black Females			
Coefficient on CVD before age 60	(1.251, 1.261)	(1.376, 1.381)	(1.502, 1.532)
Coefficient on CVD after age 60	(0.952, 1.016)	(0.974, 1.023)	(0.926, 0.975)
Coefficient on Cancer before age 60	(1.146, 1.171)	(1.185, 1.190)	(1.275, 1.280)
Coefficient on Cancer after age 60	(0.939, 1.018)	(0.955, 1.029)	(0.933, 1.012)
Function Value	−0.001101	−0.004422	−0.006474

TABLE 7B: Marginal Identified Regions (break at age 65)

	1970–1980	1970–1990	1970–2000
White Males			
Coefficient on CVD before age 65	$(1.145, 1.150)$	$(1.338, 1.348)$	$(1.433, 1.444)$
Coefficient on CVD after age 65	$(0.925, 0.999)$	$(0.936, 1.015)$	$(0.913, 0.992)$
Coefficient on Cancer before age 65	$(1.081, 1.095)$	$(1.100, 1.100)$	$(1.183, 1.200)$
Coefficient on Cancer after age 65	$(0.923, 1.002)$	$(0.938, 1.017)$	$(0.936, 1.015)$
Function Value	-0.001786	-0.006560	-0.006840
White Females			
Coefficient on CVD before age 65	$(1.146, 1.146)$	$(1.201, 1.221)$	$(1.273, 1.273)$
Coefficient on CVD after age 65	$(0.932, 1.006)$	$(0.969, 1.043)$	$(0.958, 1.032)$
Coefficient on Cancer before age 65	$(1.091, 1.099)$	$(1.145, 1.149)$	$(1.232, 1.237)$
Coefficient on Cancer after age 65	$(0.938, 1.017)$	$(0.930, 1.004)$	$(0.956, 1.016)$
Function Value	-0.001405	-0.003768	-0.004637
Black Males			
Coefficient on CVD before age 65	$(1.145, 1.145)$	$(1.233, 1.249)$	$(1.388, 1.388)$
Coefficient on CVD after age 65	$(1.010, 1.024)$	$(0.931, 1.010)$	$(0.932, 1.011)$
Coefficient on Cancer before age 65	$(1.001, 1.031)$	$(1.001, 1.026)$	$(1.078, 1.098)$
Coefficient on Cancer after age 65	$(0.917, 0.996)$	$(0.934, 1.013)$	$(0.937, 1.016)$
Function Value	-0.000768	-0.006815	-0.010775
Black Females			
Coefficient on CVD before age 65	$(1.182, 1.192)$	$(1.254, 1.259)$	$(1.388, 1.388)$
Coefficient on CVD after age 65	$(0.932, 1.011)$	$(0.989, 1.033)$	$(0.882, 0.961)$
Coefficient on Cancer before age 65	$(1.102, 1.107)$	$(1.147, 1.147)$	$(1.169, 1.199)$
Coefficient on Cancer after age 65	$(0.931, 1.011)$	$(0.957, 1.016)$	$(0.987, 1.066)$
Function Value	-0.005226	-0.012324	-0.013312

TABLE 8: Marginal Identified Regions Excluding Lung Cancer

	1970–1980	1970–1990	1970–2000
White Males			
Coefficient on CVD	$(1.126, 1.129)$	$(1.295, 1.296)$	$(1.392, 1.399)$
Coef. on Cancer (excl. lung)	$(1.091, 1.093)$	$(1.039, 1.045)$	$(1.236, 1.249)$
White Females			
Coefficient on CVD	$(1.091, 1.093)$	$(1.201, 1.206)$	$(1.267, 1.269)$
Coef. on Cancer (excl. lung)	$(1.126, 1.129)$	$(1.239, 1.249)$	$(1.455, 1.458)$
Black Males			
Coefficient on CVD	$(1.126, 1.129)$	$(1.202, 1.206)$	$(1.334, 1.346)$
Coef. on Cancer (excl. lung)	$(1.112, 1.115)$	$(1.201, 1.205)$	$(1.118, 1.119)$
Black Females			
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 9: Marginal Identified Regions (only Cancer multiplicative)

	1970–1980	1970–1990	1970–2000
White Males			
Coefficient on Cancer	$(0.520, 2.186)$	$(0.602, 2.124)$	$(0.654, 2.124)$
White Females			
Coefficient on Cancer	$(0.802, 1.610)$	$(0.890, 1.646)$	$(1.002, 1.698)$
Black Males			
Coefficient on Cancer	$(0.449, 2.356)$	$(0.484, 2.200)$	$(0.550, 2.332)$
Black Females			
Coefficient on Cancer	$(0.556, 2.284)$	$(0.644, 2.230)$	$(0.702, 2.332)$

TABLE 10: Marginal Identified Regions Starting at Age 40

	1970–1980	1970–1990	1970–2000
White Males			
Coefficient on CVD	$(1.126, 1.129)$	$(1.295, 1.296)$	$(1.389, 1.391)$
Coefficient on Cancer	$(1.001, 1.029)$	$(1.020, 1.035)$	$(1.134, 1.153)$
Function	-0.019	-0.036	-0.037
White Females			
Coefficient on CVD	$(1.091, 1.093)$	$(1.158, 1.160)$	$(1.236, 1.238)$
Coefficient on Cancer	$(1.091, 1.093)$	$(1.154, 1.160)$	$(1.201, 1.206)$
Function	-0.009	-0.016	-0.021
Black Males			
Coefficient on CVD	$(1.126, 1.129)$	$(1.201, 1.206)$	$(1.334, 1.346)$
Coefficient on Cancer	$(1.030, 1.034)$	$(1.063, 1.066)$	$(1.072, 1.074)$
Function	-0.005	-0.029	-0.039
Black Females			
Coefficient on CVD	$(1.158, 1.160)$	$(1.231, 1.235)$	$(1.334, 1.346)$
Coefficient on Cancer	$(1.096, 1.096)$	$(1.167, 1.172)$	$(1.160, 1.160)$
Function	-0.029	-0.035	-0.050

Figure 1: Trends in age–adjusted mortality 1950–2000 (all persons)

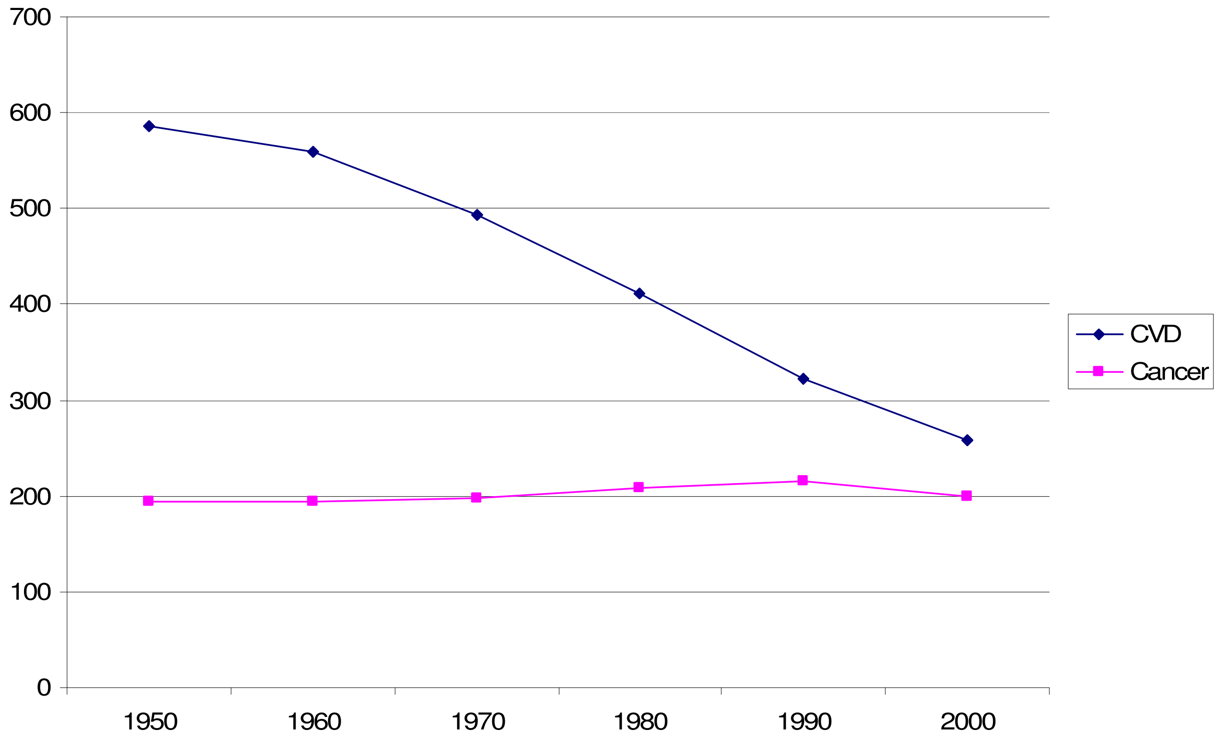

Figure 2: Hazard Rates for the Cancer

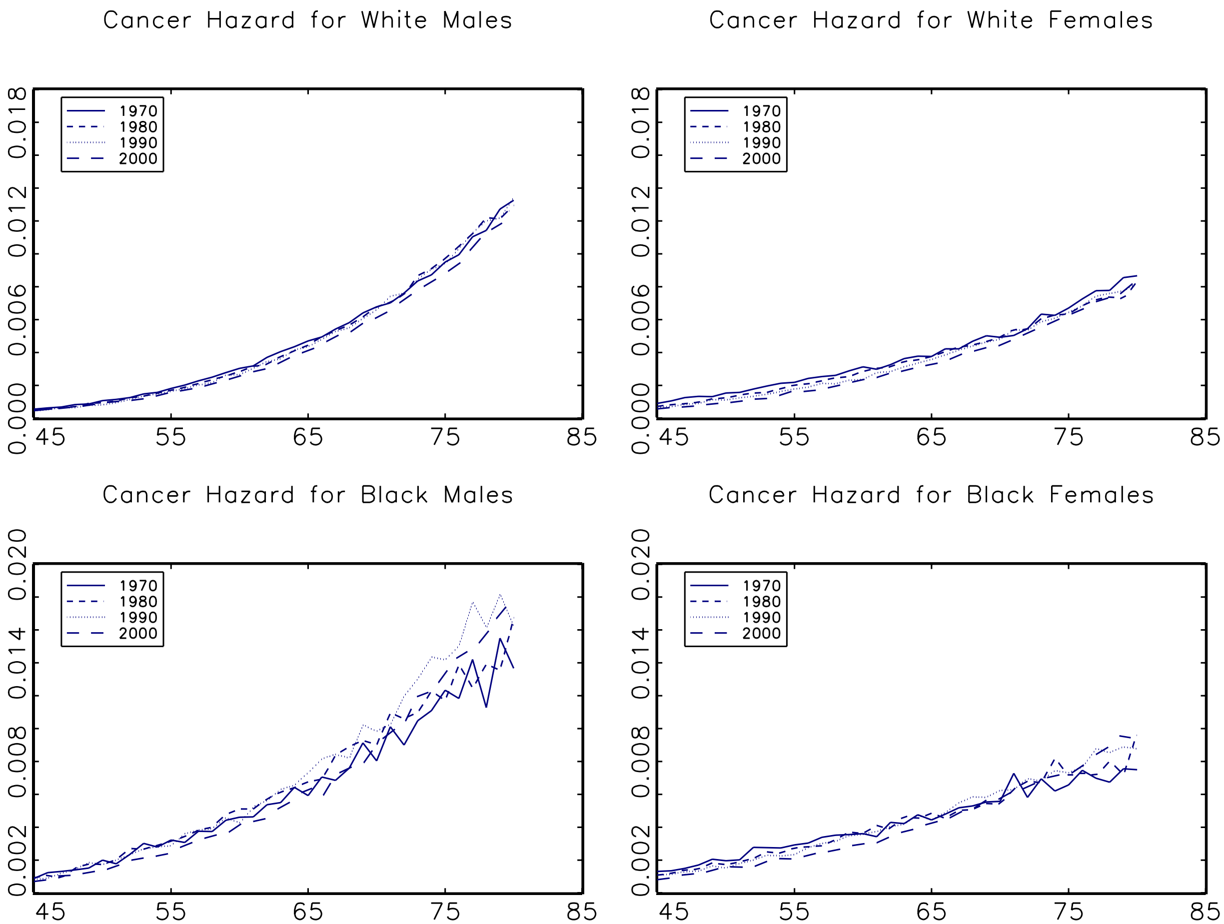

Figure 3: Hazard Rates for the CVD

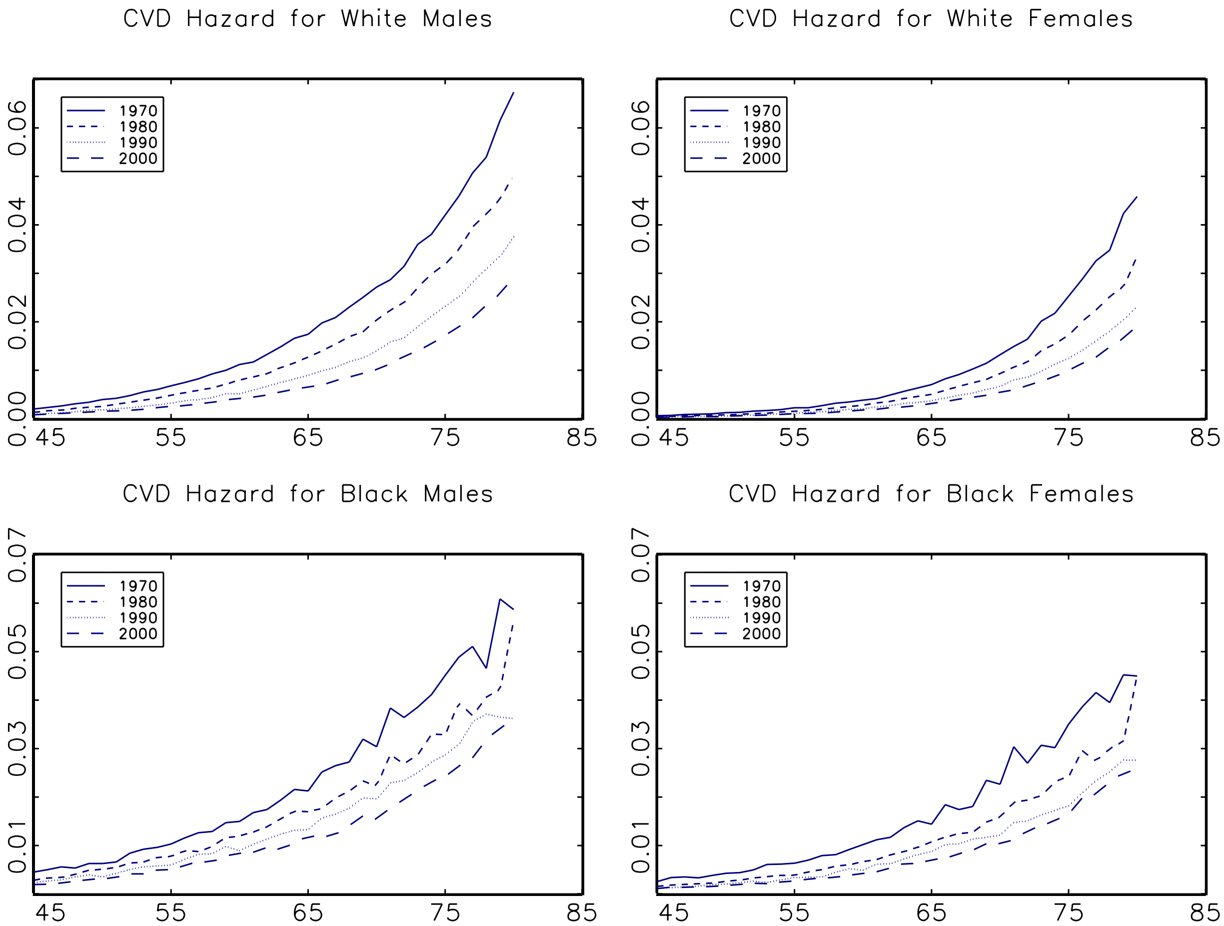

Figure 4: Peterson Bounds on the Cancer Survivor Functions in 1970 and 2000

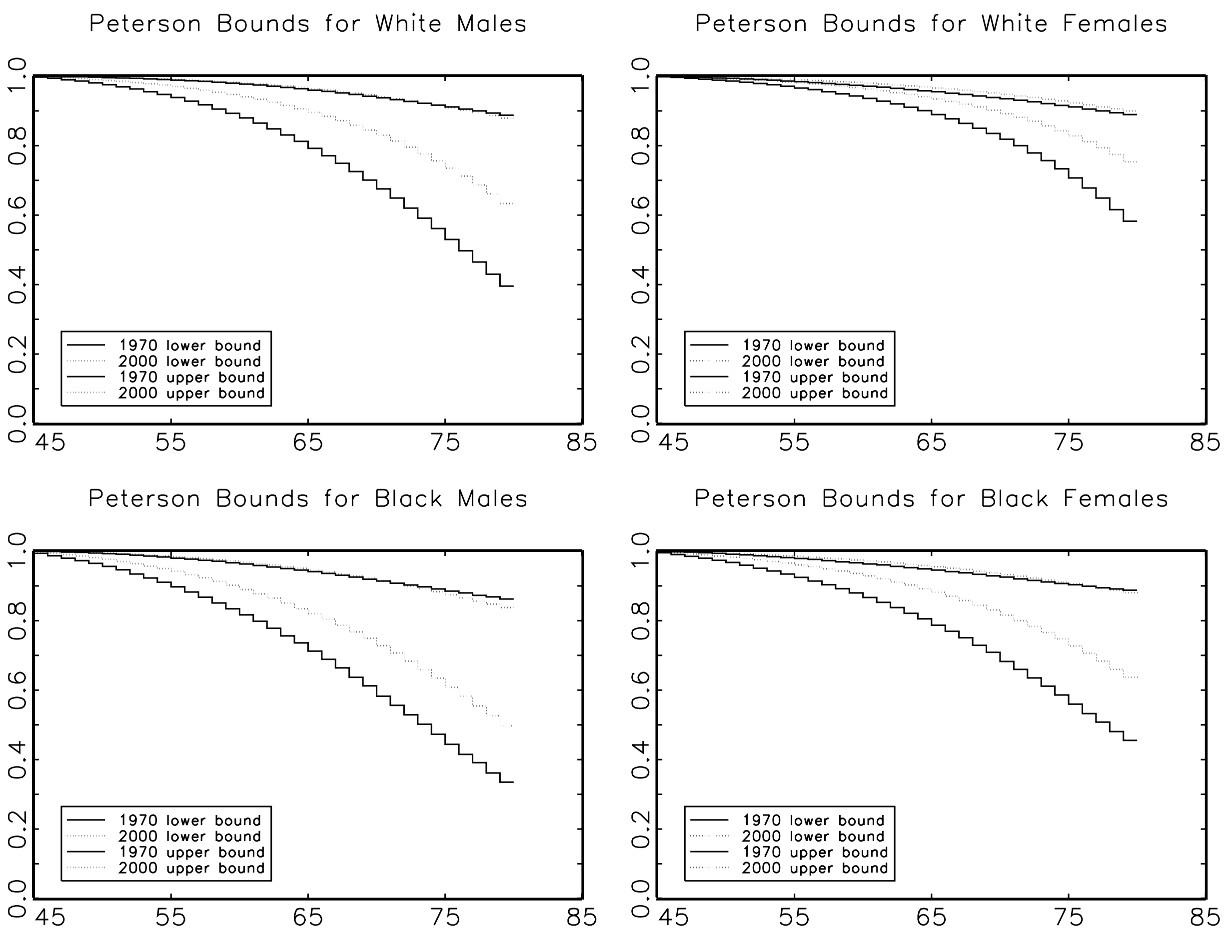

Figure 5: Illustration of Points of Support

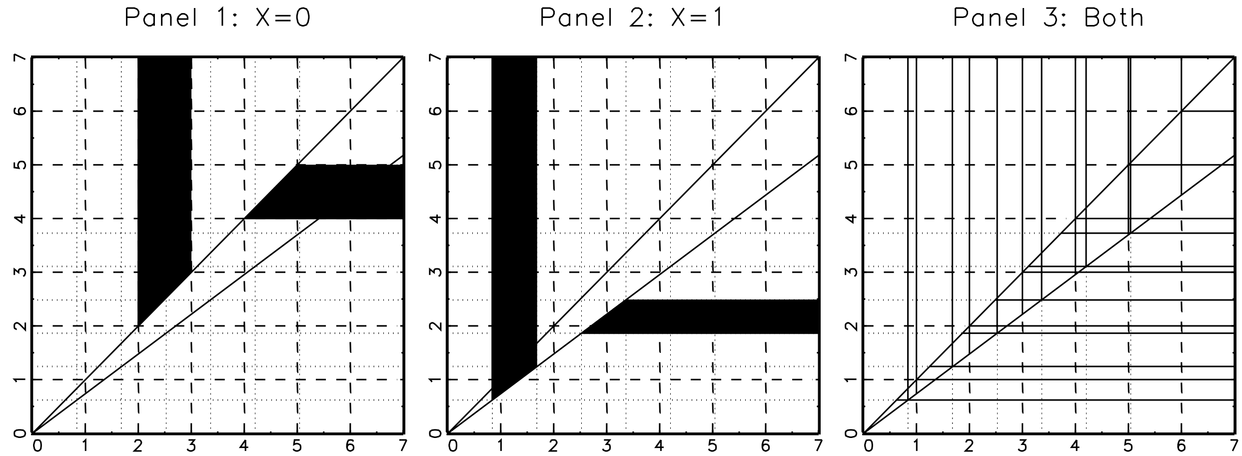

Figure 6: Allowing for Kinks: CVD

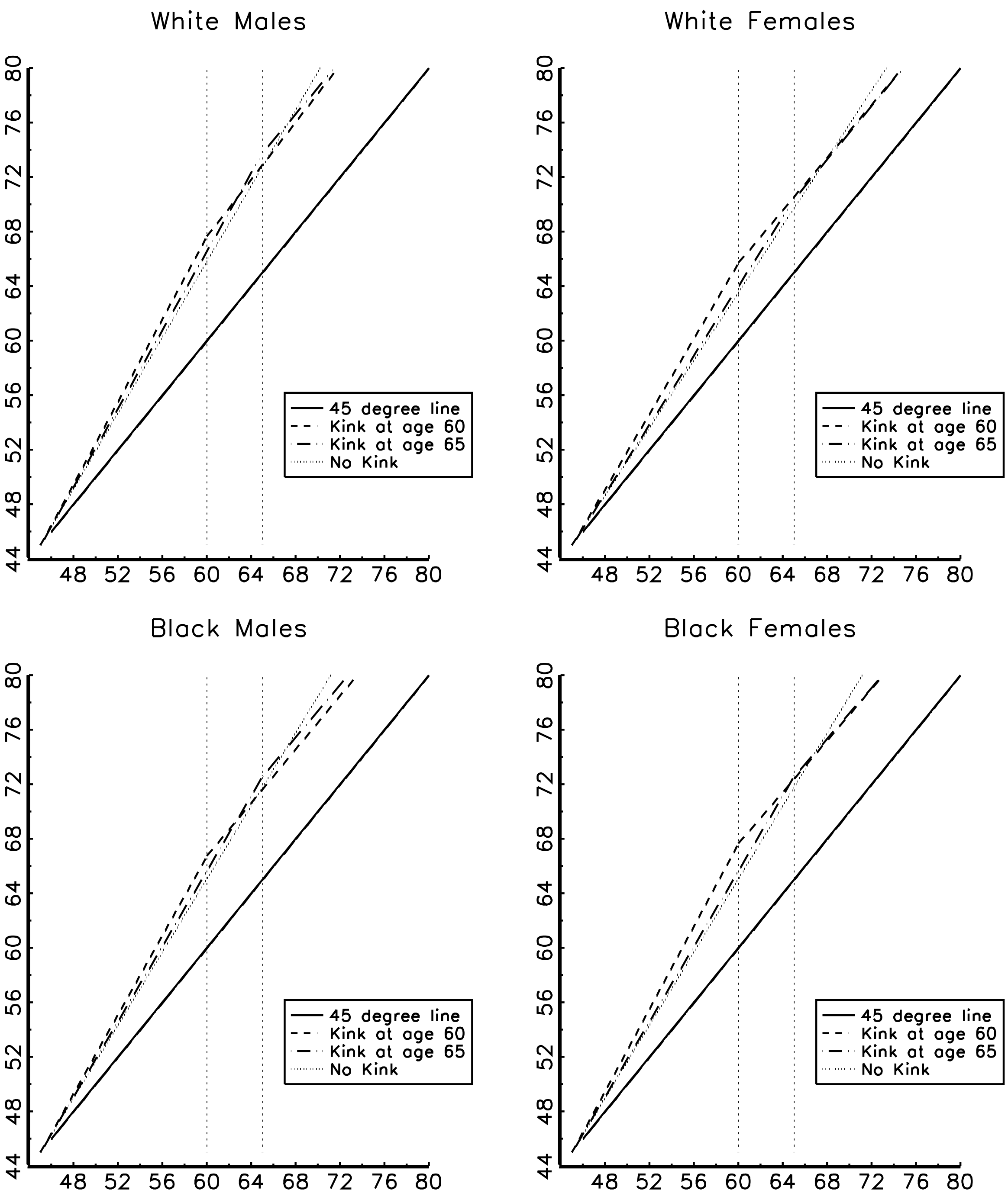

Figure 7: Allowing for Kinks: Cancer

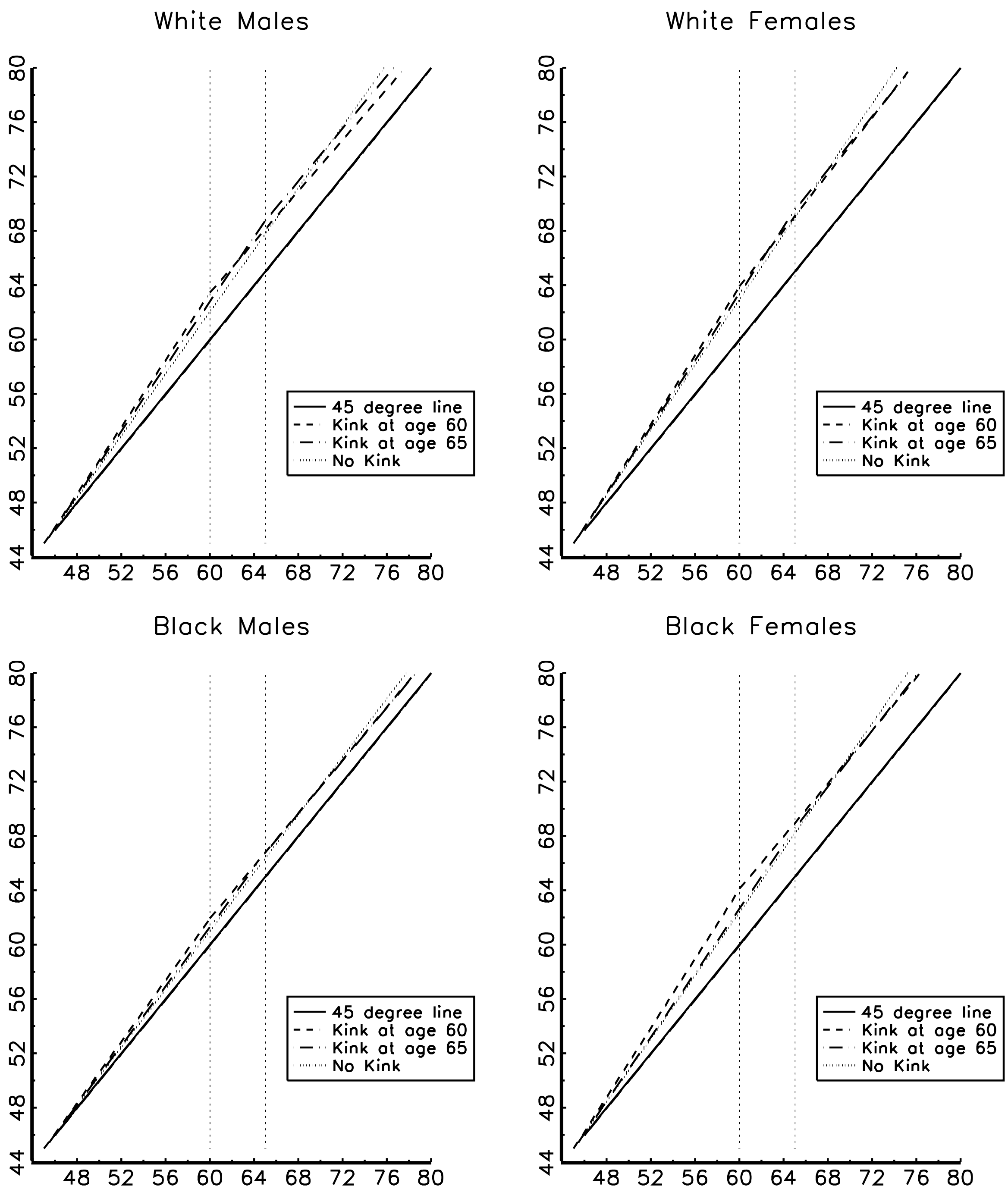

Figure 8: CVD Death Rates by Age in 1970

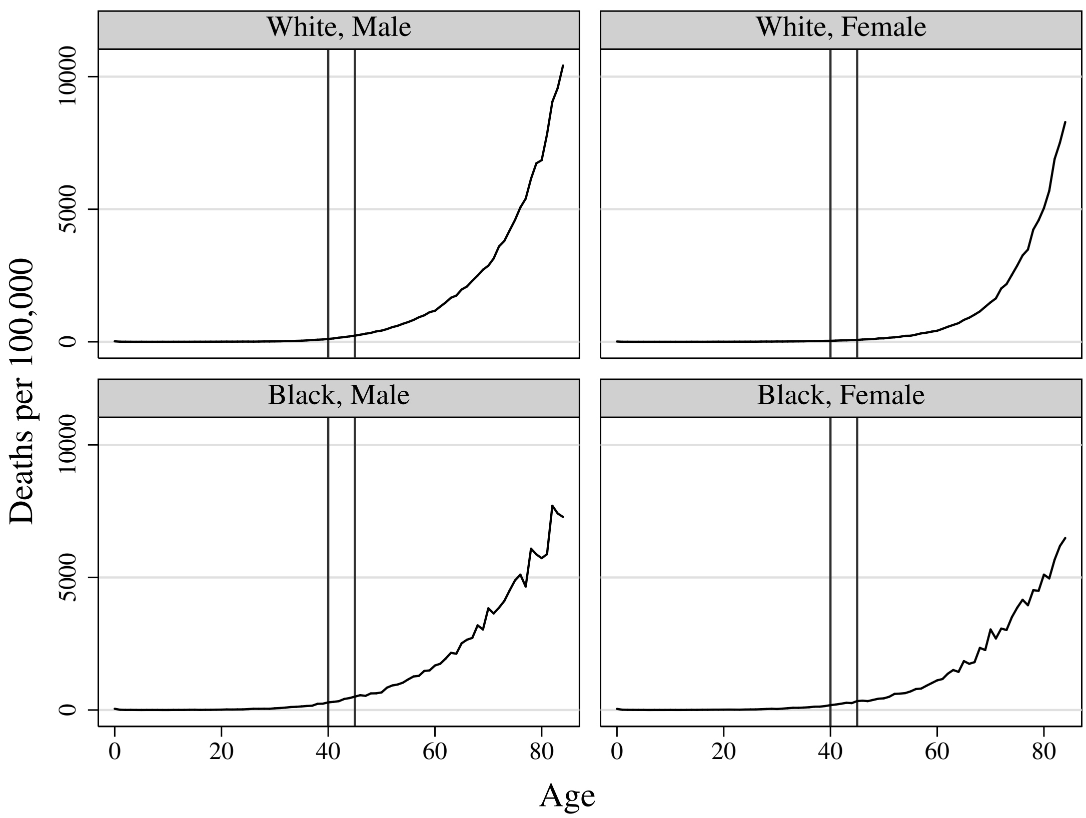

Figure 9: Cancer Death Rates by Age in 1970