CAUSE-SPECIFIC FAILURE PROBABILITY WITH COVARIATE EFFECTS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAII IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

BIOMEDICAL SCIENCES
(BIOSTATISTICS-EPIDEMIOLOGY)

MAY 1996

By

Kwisook Choi

Dissertation Committee:

Chin Sik Chung, Chairperson
Chai B. Park
Laurence N. Kolonel
Loic Le Marchand
Ming P. Mi

This dissertation is dedicated to my family.

ACKNOWLEDGMENTS

I would like to thank the dissertation committee members for their valuable suggestions and being helpful for this dissertation. I especially thank Dr. Chin Sik Chung for his valuable advice and guidance for over the years. His encouragement and precious suggestions in data analysis are invaluable for the completion of this study. I thank Dr. Chai Bin Park for introducing the theory of survival analysis, which is the basis of this dissertation; Dr. Laurence Kolonel for allowing me to access to the data of prostate cancer cases in Hawaii, which motivated me for this study; Dr. Loic Le Marchand for his valuable encouragement in the course of this study; Dr. Ming-Pi Mi for his guidance in simulation.

I also wish to acknowledge Ms. Lynne Wilkens for her critical reading of this dissertation and for valuable suggestions throughout this study; staff in the Cancer Research Center of Hawaii for being helpful; Mr. Randi Chen, Dr. Lei Li, and Mr. Sehyun Kim, for their valuable discussions.

I also thank the Department of Public Health Sciences, University of Hawaii-Manoa for providing me with the opportunity for my graduate work.

ABSTRACT

When infectious diseases were the main killers, elimination of their effects on mortality rates was possible. However, in modern society in which chronic diseases are major causes of death, elimination of disease effect is no longer relevant in estimating mortality rates.

The cause-specific failure probability (CSFP) under the unrealistic assumption of elimination of other causes (net probability) is always larger than that under the practical situation where all other causes act simultaneously (crude probability), given competing risks. The proportional hazards model is fit to estimate the CSFP when covariate effects are considered. Fitting the model is performed simply by treating the patients who fail due to other causes as censored observations. When the coefficient of a covariate is positively related to the hazards of other causes, the estimate of net probability is increased over that of crude probability according to the increasing value of the covariate. The more the association of a covariate is related to other causes, the larger the difference is.

However, the interpretation is complicated because the probability is related to one or more of the cause-specific hazards, and these hazards are also influenced by covariates. If the proportionality assumption of Cox's model is violated for an explanatory variable, stratification for the variable is desirable and CSFP is estimated in each stratum. If the proportional hazards model fits selected causes, strata, or time zones, a combination of non-parametric and semiparametric hazard and survival functions can be used to estimate the CSFP. An illustrative example is given for prostate cancer patients in Hawaii.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

What is the probability of dying due to prostate cancer for a patient with prostate cancer in a certain period of time? In this case, competing risk should be considered because the patient will eventually die due to either prostate cancer (cause of interest, or cause 1) or other diseases (other causes, extraneous causes, or cause 2) rather than prostate cancer.

The classical problem of competing risks analysis was the estimation of failure rate for some cause with the removal of some or all other causes (1-8). A historical example, reported in Karn (9), dates back to 1760, when Daniel Bernoulli was interested in estimating mortality rates if smallpox was eradicated (5). When infectious diseases were the main killers, it was natural to think of eliminating one or more causes of death by eradicating the sources of infection, or prevention of fatal development of the disease by vaccination (4,8). It was also reasonable to assume that the force of mortality for a given cause is the

same in the presence as in the absence of other causes. In that situation, $1-\exp\{-\int_0^t h_j(u)du\}$ (where j is cause, h_j is the $j-th$ cause-specific hazard, $j=1,\ldots,m$, and u is time) is an appropriate way to define a cause-specific failure probability (CSFP) and estimated by deaths due to other causes as censored observations.

At present, most infectious diseases are rather effectively controlled. Chronic diseases, associated with aging have now become the major causes of death (4). It is difficult to conceive, in the strict sense, of the elimination of a chronic disease as a cause of death, because the failure rate function for a specific failure type may be affected in a variety of ways by "removal" of other failure types. "Removal" itself may involve a variety of mechanisms within the remaining failure types. For these reasons, the estimation of failure rates for certain types of failure given the "removal" of some or all other failure types is not practical (2). Therefore, methods for the study of competing risk analysis have been developed under the practical situation of other causes acting

simultaneously without imposing the independent competing risks assumption rather than under the impractical assumption of elimination of other causes.

Methods for the analysis of cause-specific failure data in a clinical or an epidemiological setting have been developed by Crowley and Hu (10), Holt (11), Prentice et al. (1), and Kalbfleisch and Prentice (2) as a natural extension of the methodology used in the construction of multiple-decrement life tables (4). Under the most general assumption of no parametric restrictions placed on the underlying cause-specific hazards, estimation of the particular cause-specific hazard and tests of the covariate effects on that hazard are performed simply by treating the patients who fail from extraneous causes as censored observations (12). Cause-specific hazard functions are the basic estimable quantities in the competing risks framework (1). However, estimation of cause-specific failure probability and regarding other causes as censored requires the unrealistic assumption that extraneous causes are eliminated and do not affect the cause of interest. It means that the "usual" cause-specific cumulative failure

function ($1-\exp\{-\int_0^t h_j(u)du\}$) is impractical way to define a cause-specific failure probability. But in the clinical oncology and epidemiology literature, it is still quite common to see cause-specific failure probability defined as the formula, $1-\exp\{-\int_0^t h_j(u)du\}$ (12). To be at risk of failure at a particular time *t*, one must first survive all causes of failure until that time. The formula, $1-\exp\{-\int_0^t h_j(u)du\}$, does not account for more than one cause of failure (net probability) and thus will produce an estimate of the probability of being at risk of failure at time *t* larger than the actual probability. This leads to an estimate of the net probability larger than the practical CSFP.

The likelihood of failing from a particular cause at time *t* is the product of the overall survival function times the cause-specific hazard at time *t*. The integral of this function over (0,t] ($\int_0^t h_j(u)S(u)du$) is the practical cause-specific failure probability, given competing risks (crude probability). Here each cause-specific failure probability is defined as a function of all the cause-specific hazards

via the joint survival functions. Therefore, the interpretation of the cause-specific probability of failure is somewhat complicated. Two groups of patients with the same marginal risk of death due to cause 1 may have different cause-specific failure probabilities due to differences in other cause-specific hazards (12).

If subjects are not homogeneous in characteristics, e.g., different age, race, behavior,..., as in prostate cancer patients in Hawaii, covariate effects on the hazard should be considered. The proportional hazards model of Cox (13) can be fit to account for the covariate effects on the hazard. The Cox's hazards model is applied in the same way for competing risks data as for data with one cause of failure by treating extraneous causes as censored at their times of failure. The estimate for the cause-specific failure probability considering covariate effects based on net probability (1- exp{- cause-specific cumulative hazard with covariate effects}) is not an practical estimate, given competing risks (14). The function of the cause-specific failure probability with the covariate effects based on the crude probability was reported by Kalbfleisch and Prentice

(2). They also provided a method for checking the proportionality assumption for the fit of Cox's model in the presence of competing risks. Some methods to estimate CSFP with the covariate effects are derived from semiparametric estimates of the absolute cause-specific risk provided by Benichou and Gail(15). The formula will be explained in Chapter 3. The interpretation is complicated because the probability is related to one or more of the cause-specific hazards, and these hazards are influenced by covariates. This will be discussed in Chapter 5. In this report, it is assumed that observations are available on the failure time of n individuals taken to be independent.

The main focus of this study is on cause-specific failure probability with covariate effects. To help understand this concept, functions of hazard and cause-specific failure probability will be explained for the subjects assumed to be homogeneous in characteristics (without covariates) at first, and then both functions will be explained for the subjects different in characteristics (with covariates).

The purpose of this study is to investigate the appropriate formulation for the probability of failure in the presence of competing risks and covariate effects.

Specifically, the objectives of this study include:

[1] To examine the difference between the estimate of net probability and that of the crude probability for simulated data and prostate cancer patients in Hawaii assumed to be homogeneous in characteristics.

[2] To examine the difference between the estimate of net probability and that of crude probability with covariate effects for prostate cancer patients in Hawaii.

[3] To clarify the concepts of maximum likelihood estimates of cause-specific hazard functions and cause-specific failure probabilities under the settings of homogeneous subjects and heterogeneous subjects.

[4] To provide alternative methods to estimate cause-specific failure probability with covariate effects when the proportionality assumption for covariates is violated.

[5] To discuss some limitations in the interpretation of cause-specific failure probabilities with covariate effects.

This dissertation is organized as follows. A review of the literature on relevant issues follows immediately. Chapter 3 is devoted to the methodological development of competing risk survival analysis. Models and the estimation procedure will be explained. Chapter 4 presents the results using simulated data and real data. The comparisons of the estimates of net probability and crude probability are shown. Depending on the appropriateness of the proportional hazards models for selected causes, strata, or time zones, a combination of non-parametric and semiparametric hazard and survival functions are suggested. A summary and discussion of the major findings are given in Chapter 5.

CHAPTER 2

REVIEW OF LITERATURE

This chapter is divided into three sections. Section 2.1 introduces studies of competing risks including historical examples and some important studies related to this dissertation. Sections 2.2 and 2.3 provide basic theory for survival analysis and the proportional hazards model. Specific theories for competing risks analysis will be introduced in the next chapter.

2.1 Studies of Competing Risks

2.1.1 Origin of Competing Risks

The beginnings of the theory of competing risks can be traced to a memoir read in 1760 by Daniel Bernoulli before the French Academy of Science (5). Trained in medicine as well as mathematics, Bernoulli arrived at a mathematical solution, under certain assumptions, to the following question of great interest at the time: If in a given population smallpox could be eradicated, what would be the

effect on the population mortality structure at different ages? Using Halley's Breslau life table of 1693, Bernoulli actually constructed a hypothetical life table corresponding to the elimination of smallpox. A key assumption was, as he recognized, that individuals saved from smallpox were subject to other causes of death in exactly the same manner as the rest of the population. This independence assumption would not hold if smallpox tended to sweep away only the weakest members of the population, a supposition Bernoulli finds reason to dismiss.

Bernoulli assumed that in a year smallpox attacks one out of every n who have not had the disease and that one out of m attacked dies. Although he spent considerable efforts in justifying these assumptions and in estimating m and n on the basis of available data, the assumptions were unnecessarily restrictive (5). His severest critic D'Alembert was unduly harsh and frequently in the wrong but he did provide a modification of Bernoulli's approach which has the required flexibility. At that time, the focus of interest was on infectious diseases, such as smallpox, measles, typhus, etc. In this context, elimination of a

cause has real natural meanings: removal of the source of infection, or prevention of fatal development of the disease by vaccination (4,8). Thus in 1874 William Farr (16) found it worthwhile to examine the "Effect of the extinction of any single disease on the duration of life" in a supplement to the Annual Report of the Registrar General. In that report Farr refers to earlier discussions on the topic by Daniel Bernoulli and D'Alembert.

Karn (9) in 1931 gives an interesting historical account of the works of Bernoulli and D'Alembert with a detailed review of their mathematical contributions, and those of later writers (5). This paper includes applications of the early methods to 20th century data. Surprisingly, however, Karn does not refer to Todhunter's 1865 *History* (17) and does not seem to have been aware of Makeham's (18) contribution. Todhunter, presenting some of the same early work, concentrates on the mathematics but allows himself the comment: "Since the introduction of vaccination, the memoir of Bernoulli will have no practical value." The British actuary, W. M. Makeham was considering theoretical problems involved in preparing life tables for

populations subject to multiple sources of decrement; he drew upon and simplified the earlier work so that the Bernoulli-D'Alembert-Makeham method provides the basis for current actuarial methods of handling this problem (3,6).

Chronic diseases have now become the major causes of death. Development of a disease is usually a complicated physiological process, involving metabolic errors (failures) of various components (4). Some of these components might be common to two or more diseases, others are different. As long as we cannot identify the factors that take part in the disease process, we cannot determine the effect of elimination of one cause of death on the mortality pattern from other causes.

Therefore, studies on CSFP are appropriate under the practical situation of other causes acting simultaneously without imposing the independent competing risks assumption rather than under the impractical assumption of elimination of other causes.

2.1.2 Cornfield's Empirical Effects

Cornfield introduced some terminology, pure probability and mixed probability, in his article "The estimation of the probability of developing a disease in the presence of competing risks" in 1957 (6). He noted that the mixed and pure probabilities are identical to what Fix and Neyman (19,20) have called crude and net probabilities, respectively, in a problem in long-term survival involving competing risks. (These probabilities are defined in the next section.) He introduced how to estimate pure probability mathematically and mentioned that the basic analytic apparatus necessary for making the estimates was developed by Bernoulli and D'Alembert as part of a consideration of the effect of eliminating smallpox as a cause of death on the probability of surviving to a given age. He gave a detailed concept of mixed probability but no mathematical estimation of it. He used projected empirical data to get an estimate of mixed probability and compared it with pure probability and called the difference of two probabilities an empirical effect. He said, "With respect to actual knowledge of the magnitude of possible empirical

effects of competing risks we seem to have made no advance beyond Bernoulli. One way of expanding our knowledge and insight in this area is by means of the kind of experimental epidemiology". It seems that he thought that empirical effects could not be adequately estimated by mathematics alone.

2.1.3 The Crude, Net, and Partial Crude Probability

Chiang mentioned three types of probability of death and their relationships in study of competing risks (3).

1. The crude probability. The probability of death from a specific cause in the presence of all other risks acting in a population, or

$Q_{i\delta} = \Pr\{$an individual alive at time x_i will die in the interval$(x_i,\ x_{i+1})$ from cause R_δ in the presence of all other risks in the population$\}$

2. The net probability. The probability of death if a specific risk is the only risk in effect in the population or, conversely, the probability of death if a specific risk is eliminated from the population.

$q_{i\delta} = \Pr\{$an individual alive at x_i will die in the interval (x_i, x_{i+1}) if R_δ is the only risk acting on the population$\}$;

$q_{i.\delta} = \Pr\{$an individual alive at x_i will die in the interval (x_i, x_{i+1}) if R_δ is eliminated as a risk of death$\}$.

3. The partial crude probability. The probability of death from a specific cause when another risk (or risks) is eliminated from the population.

$Q_{i\delta.1} = \Pr\{$an individual alive at x_i will die in the interval (x_i, x_{i+1}) from R_δ if R_1 is eliminated as a risk of death$\}$;

$Q_{i\delta.12} = \Pr\{$an individual alive at x_i will die in the interval (x_i, x_{i+1}) from R_δ if R_1 and R_2 are eliminated as risks of death$\}$.

When the cause of death is not specified,

$p_i = \Pr\{$an individual alive at x_i will survive the interval $(x_i, x_{i+1})\}$ and

$q_i = \Pr\{$an individual alive at x_i will die in the interval $(x_i, x_{i+1})\}$, with $p_i + q_i = 1$.

The terms "risk" and "cause" may refer to the same condition, but are distinguished here by their position in time relative to the occurrence of death. Prior to death the condition referred to is a risk; after death the same condition is the cause.

Because of the absence of competing risks, the net probability is always greater than the corresponding crude probability. In the human populations, the net and partial crude probabilities cannot be estimated directly, but only through their relations with the crude probability. The study of these relations is part of the problem of "competing risks," or "multiple decrement."

2.1.4 Study by Gaynor et al.

The article entitled "On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data" was published in the Journal of the American Statistical Association in June 1993 (12). The authors, Gaynor et al., pointed out that it is still quite common to see the cause-specific failure probability incorrectly estimated as the one minus the Kaplan-Meier (1-

KM) estimator treating patients who fail of extraneous causes as censored observations. They presented examples showing the extent of bias applying the 1-KM approach in clinical oncology data, and concluded that the bias can be quite large if there is little censoring or if a large percentage of patients fail from extraneous causes prior to the occurrence of failures from the cause of interest.

2.1.5 Other Studies of Competing Risks

The actuarial approach to competing problems, in the absence of regressor variables, has been discussed by many authors (2). For example, besides Cornfield (6) and Chiang (3), Seal (1954), Elveback (21), Kimball (22), Berkson and Elveback (23), Pike (24), and Hoel (25) discuss models and statistical methods for the estimation of crude, partial crude, and net probabilities. An assumption of independence between latent failure times pervades most of this work though Kimball (22) considers a somewhat different, but equally strong assumption. Gail (26) has provided an excellent review of literature on competing risk analysis.

Aalen (27) has considered some formal properties of nonparametric estimators for the multiple decrement model.

Many authors have considered parametric models for competing risk data. Marshall and Olkin (28), Moeschberger and David (29), and Moeschberger (30) utilize exponential and Weibull models. The Marshall and Olkin bivariate models include a positive probability that the latent failure times (T_1, T_2) are equal. Nadas (31) has used a bivariate normal model and Hoel (25) fitted an independent risks Makeham-Gompertz model. A historical perspective is given by Seal (32). The nonparametric estimation of cause-specific failure probability or cumulative incidence function was studied by Aalen (33), Elandt-Johnson and Johnson (4), Lawless (34), Namboodiri (8), Gray (35), Benichou and Gail (15), Babu, Rao, and Rao (36), and Huang and Wang (37).

The proportional cause-specific hazards function model has been utilized by Holt (11), Farewell (38), Prentice and Breslow (39), Prentice et al. (1) , Kalbfleisch and Prentice (2), Cox and Oakes (7), Kay (40), Larson and Dinse (41), Kuk (42), and Lunn and McNeil (43). Two approaches have been followed for analyzing cause-specific survival data with

covariate effects. The first, described by Kalbfleisch and Prentice (2) and used, for example, by Kay (40), simply involves fitting models separately for each type of failure in turn, treating other failure types as censored data. This method is also used for this study. The other approach, which has been used by various authors including Larson and Dinse (41) and Kuk (42), involves fitting more complex models incorporating the different failure types. Lunn and McNeil (43) gave for the joint estimation of parameters in Cox's proportional hazards regression model (13) using a data duplication method.

2.2 Functions of Survival Time and Product-Limit Estimates

Competing risk survival analysis is the generalization of ordinary survival analysis (44). The modern statistical analysis of survival data was catalyzed by the presentation in 1958 by Kaplan & Meier (45) of some of the statistical properties of the product limit estimator, a preexisting

variation of the life table in which the necessity for grouping the data was removed (14).

2.2.1 General functions of Survival Time

Let T be a non-negative random variable representing the time to failure. The distribution of T can be characterized by following (2-4,7-8,14,34,46,47).

The survival function $S(t)$ is defined as the probability that an individual survives longer than t:

$$S(t) = p(T > t). \tag{2.1.1}$$

From the definition of the cumulative distribution function $F(t)$ of T,

$$S(t) = 1 - P(\textit{an individual fails before or equal to time } t)$$

$$= 1 - F(t) \tag{2.1.2}$$

The probability density function $f(t)$, the derivative of the cumulative distribution, is the probability of failure in a small interval per unit time:

$$f(t) = -S'(t) = \lim_{\Delta \to 0+} \frac{S(t) - S(t+\Delta)}{\Delta} \tag{2.1.3}$$

The cumulative distribution function is,

$$F(t) = \int_0^t f(x)dx \tag{2.1.4}$$

The hazard function $h(t)$ is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the interval,

$$h(t) = \frac{f(t)}{S(t)} \tag{2.1.5}$$

$$= -\frac{S'(t)}{S(t)}$$

$$= -\frac{d}{dt}\log S(t) \tag{2.1.6}$$

The cumulative hazard function is defined as

$$H(t) = \int_0^t h(x)dx\ . \tag{2.1.7}$$

From (2.1.6) and (2.1.7),

$$H(t) = -\log S(t)$$

or

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(x)dx\right] \tag{2.1.8}$$

From (2.1.5) and (2.1.8),

$$f(t) = h(t)S(t) = h(t)\exp[-H(t)] \tag{2.1.9}$$

The cumulative distribution function or probability of failure is

$$F(t) = \int_0^t f(x)dx = \int_0^t h(x)S(x)dx \qquad (2.1.10)$$

2.2.2 Product-Limit Estimates of Survivorship Function

Intuitively, from the above functions of survival time, an estimate of the hazard or risk $h(t_i)$ of dying at t_i, given survival to that time, is given by the fraction of those at risk who do die at t_i; $\frac{d_i}{n_i}$ (where d_i = deaths at time t_i, and n_i = the number at risk at t_i) (2-4,7-8,14,34,46,47). Thus, a simple estimator of the cumulative hazard function $H(t)$ is given at t_i by $\sum_{k=1}^{i} \frac{d_k}{n_k}$, with interpolation, perhaps by stair-step, between true survival times. This estimator translates into an estimator for $S(t_i)$ as

$$\tilde{S}(t_i) = \exp\left[-\sum_{k=1}^{i} \frac{d_k}{n_k}\right], \qquad (2.1.11)$$

with estimated standard error

$$\left[\sum_{k=1}^{i} \frac{d_k}{n_k(n_k - d_k)}\right]^{1/2}. \qquad (2.1.12)$$

Historically, estimators of $S(t)$ were derived using products of conditional probabilities of surviving intervals, using the fact that

$$S(t_i) = \prod p(X > t_k | X > t_{k-1}) = \prod_{k=1}^{i} \frac{S(t_k)}{S(t_{k-1})} . \qquad (2.1.13)$$

If the intervals are chosen, using successive distinct survival times, the conditional probabilities can be estimated by the fraction of those at risk surviving each interval, $\frac{(n_k - d_k)}{n_k}$, giving the product limit or Kaplan-Meier (45) estimator of $S(t_i)$:

$$\hat{S}(t_i) = \prod_{k=1}^{i} (1 - \frac{d_k}{n_k}) . \qquad (2.1.14)$$

For t in (t_i, t_{i+1}), the estimator is usually taken to be constant, and is thus a stair-step with jumps at the true survival times. The standard error at t is estimated by Greenwood's formula (48):

$$\hat{S}(t) \left[\sum_{k: t_k \le t} \frac{d_k}{n_k (n_k - d_k)} \right]^{1/2} . \qquad (2.1.15)$$

When the d_i are small, the estimators (2.1.11) and (2.1.14) will be close, as $-\log(1 - \frac{d_i}{n_i}) \approx \frac{d_i}{n_i}$, but not identical; their

statistical properties are similar. The product limit form of Eq.(2.1.14) is the most common, partly for historical reasons and partly because it reduces to the usual empirical distribution function (a stair-step with jumps of size $\frac{1}{N}$ at each event) when there is no censoring.

2.3 The Proportional Hazards Model

One of the most widely used set of statistical techniques is regression, in which a response variable Y is related to a (column) vector of possible explanatory variables z through an assumed relationship between the expectation of Y and a function of z involving an unknown (column) vector of parameters β (14). The parameters are generally estimated by least-squares, which is most appropriate when Y is normally distributed but valid in any case, and a predictive equation formed; hypotheses about the relationship of Y and z translate to hypotheses about β. When the response Y is subject to censoring there is no immediately obvious way to extend this methodology. The optimal solution if Y (or a function such as log Y) is

assumed to be normally distributed is through the likelihood function; this no longer leads to least-squares but to a solution that must be found by iterative calculations and which is not valid when Y is not normal. The survival time is not normally distributed and subject to censoring. There are now generalizations of least-squares available (49,50,51) and a general method for regression based on ranks has also been presented, but by far the most widespread approach, due to Cox (13), focuses not on expectation or some other measure of the center of a distribution but on the hazard function. The beginnings of this idea can be found in Feigl & Zelen (52), who assumed that the hazard depended on $\mathbf{z}$ and β (and exponential model for Y), but Cox's formulation allows an arbitrary hazard function (14).

2.3.1 The Proportional Hazards Model

Let $h(t;z)$ represent the hazard function at time t for an individual with covariates $\mathbf{z}$. The proportional hazards model specifies that

$$h(t;z) = h_0(t)\exp^{z\beta} \tag{2.2.1}$$

where β is a vector of coefficients and the $h_0(t)$ is an arbitrary unspecified base-line hazard function at time t (13,2). The model is partially non-parametric since the $h_0(t)$ is not specified but the influence of the covariates is modeled parametrically as a multiplicative effect on the hazard.

The conditional probability density function of T given $\boldsymbol{z}$ corresponding to (2.2.1) is

$$f(t;z) = h_0(t)e^{z\beta}\exp\left[-e^{z\beta}\int_0^t h_0(u)du\right]. \tag{2.2.2}$$

The conditional survival function for T given $\boldsymbol{z}$ is

$$S(t;z) = [S_0(t)]^{\exp(z\beta)} \tag{2.2.3}$$

where

$$S_0(t) = \exp\left[-\int_0^t h_0(u)du\right].$$

Thus the survivor function of t for a covariate value, $\boldsymbol{z}$, is obtained by raising the base-line survivor function $S_0(t)$ to a power.

If $h_0(\cdot)$ is arbitrary, this model is sufficiently flexible for applications. There is an important

generalization that do not substantially complicate the estimation of β. The nuisance function $h_0(t)$ can be allowed to vary in specific subsets of the data. Suppose the population is divided into r strata and that the hazard $h_j(t;z)$ in the jth stratum depends on an arbitrary shape function $h_{0j}(t)$ and can be written as

$$h_j(t,z) = h_{0j}(t)\exp(z\beta), \tag{2.2.4}$$

for $j=1,\ldots,r$. Such a generalization is useful, for instance, if some explanatory variable or variables do not appear to have a multiplicative effect on the hazard function (i.e., does not fit the proportionality assumption). The range of such variables can then be divided into strata with only the remaining regression variables contributing to the exponential factor in (2.2.4). The second important generalization allows the regression variable z to depend on the time itself. Such regression variables arise in the Stanford heart transplant example (10) where treatment group itself is "time dependent", as are certain donor recipient matching variables (2).

2.3.2 Discrete Proportional Hazards Model

A discrete analogue of the proportional hazards model can be obtained by applying the survivor function relationship (2.2.3) directly to a discrete model (2). Let the failure time T given covariates $\boldsymbol{z}$ have a discrete distribution with mass points at $0 \le x_1 < x_2 < \cdots$ and let $S_0(t)$ represent the base-line survivor function for $z = 0$. The corresponding survivor function for covariates $\boldsymbol{z}$ is

$$S(t;z) = S_0(t)^{\exp(z\beta)} \tag{2.2.5}$$

as in (2.2.3). If the hazard function corresponding to S_0 has contribution h_{0i} at x_i, then

$$S_0(t) = \prod_{i|x_i<t} (1-h_{0i})$$

and, from (2.2.5),

$$S(t;z) = \prod_{i|x_i<t} (1-h_{0i})^{\exp(z\beta)} . \tag{2.2.6}$$

The hazard at x_i for covariate $\boldsymbol{z}$ is then

$$1-(1-h_{0i})^{\exp(z\beta)} . \tag{2.2.7}$$

CHAPTER 3

METHODS

This chapter illustrates the formulations necessary for this study and gives study methods.

The formulations of the cause-specific hazard function and the cause-specific failure probability function will be explained for both homogeneous subjects and heterogeneous subjects. In homogeneous subjects, because no difference exists in characteristics of subjects, e.g., same age, same race, same behavior,..., we do not need to account for covariate effects. In heterogeneous subjects, because differences exist in characteristics of subjects, we must account for covariate effects.

3.1 Homogeneous Subjects

3.1.1 Formulations

Denote T as the failure time and J as the cause of failure for each individual. Cause-specific hazard rates (2,34) can be expressed by

$$h_j(t) = \lim_{\Delta t \to 0} \frac{\Pr[t \le T < t + \Delta t, cause = j | T \ge t]}{\Delta t}, j = 1, \ldots, m. \quad (3.1.1)$$

The function $h_j(t)$ simply gives the instantaneous failure rate from cause j at time t, in the presence of the other failure types. This is called crude hazard rate because it describes the rate of dying from cause j at time t when all causes are acting simultaneously (4). Survivor, or probability density functions can also be defined as follows for $j = 1, \ldots, m$:

$$S_j(t) = Pr(T>t, \ cause=j) \quad (3.1.2)$$

$$f_j(t) = \frac{-dS_j(t)}{dt} \quad (3.1.3)$$

Any of these sets of functions specified the joint distribution of lifetime, T, and the cause of death, J. The overall distribution of T has hazard and survivor functions and Probability density function, respectively as

$$h(t) = \sum_{j=1}^{m} h_j(t)$$

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \sum_{j=1}^{m} S_j(t) \quad (3.1.4)$$

$$f(t) = -S'(t) = \sum_{j=1}^{m} f_j(t) .$$

Cause-specific hazard, or probability density function can be formulated as

$$h_j(t) = \frac{f_j(t)}{S(t)} . \tag{3.1.5}$$

$$f_j(t) = h_j(t)S(t)$$

Cause-specific failure probability under the presence of all other causes is derived as

$$F_j(t) = \int_0^t f_j(u)du$$

$$= \int_0^t h_j(u)S(u)du \tag{3.1.6}$$

It is convenient to define pseudocumulative hazard and survivor functions related to the $h_j(t)$'s. Suppose

$$H_j(t) = \int_0^t h_j(u)du$$

$$G_j(t) = \exp[-H_j(t)] . \tag{3.1.7}$$

These have the forms of cumulative hazard and survivor functions, though they do not represent the distribution of any specific random variable.

$G_j(t)$ is the "usual" cause-specific survival function where deaths from extraneous causes are treated as censored observations. It can be regarded as the "pure" distribution that describes the time to failure of the cause of interest

if no other risks were acting to kill the patient first (9). From (3.1.4) and (3.1.7) that

$$S(t) = \prod_{j=1}^{m} G_j(t) \tag{3.1.8}$$

and

$$H(t) = -\log S(t) = \sum_{j=1}^{m} H_j(t)$$

and that the $G_j(t)$'s or $H_j(t)$'s also uniquely specify the distribution of T and J.

3.1.2 Estimations

Suppose that observations are taken on a random sample of n individuals. Assume that censoring is possible and define the usual indicator variable δ_i taking on the value 1 if T_i is observed and 0 if it is censored. The observation on individual i is either of the form (t_i, J_i, δ_i =1) or (t_i, δ_i =0), where t_i is a failure time in the first case and a censoring time in the second (34). If there is independent censoring of the type, the likelihood function of the observations is

$$L = \prod_{i=1}^{n} f_{J_i}(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \ . \qquad (3.1.9)$$

Under parametric models inference by maximum likelihood methods is straightforward, here only nonparametric estimation procedures will be discussed. When there is no censoring, estimates of the $S_j(t)$'s are

$$\hat{S}_j(t) = \frac{\textit{Number of observations with } T \geq t \textit{ and } J = j}{n} \ .$$

Interval estimates for probabilities can be based on the binomial distribution. If there is censoring, matters are more complicated: Lawless (34) developed product-limit estimates of the $G_j(t)$'s of (3.1.7) and then used these to obtain estimates of the $S_j(t)$'s. From (3.1.5) and (3.1.8), likelihood (3.1.9) can be rewritten as

$$L = \prod_{i=1}^{n} h_{J_i}(t_i)^{\delta_i} G_{J_i}(t_i)^{\delta_i} \prod_{l \neq J_i} G_l(t_i)$$

$$= \prod_{j=1}^{m} \left(\prod_{i \in D_j} g_j(t_i) \prod_{i \in \bar{D}_j} G_j(t_i) \right) \qquad (3.1.10)$$

where $g_j(t) = -G'_j(t)$ and D_j is the set of individuals observed to die from cause j. The likelihood therefore factors to give a term for each $G_j(t)$, and by proceeding as in the

derivation of the product-limit estimate, the maximum likelihood estimate of $G_j(t)$ is

$$\hat{G}_j(t) = \prod_{\{i:t(i)\le t\}} (1 - h_{ji})$$

$$= \prod_{\{i:t(i)\le t\}} (1 - \frac{d_{ji}}{n_i}) \qquad (3.1.11)$$

where n_i is the number of individuals at risk just prior to t_i and d_{ji} is the number of individuals dying from cause j at t_i. An alternative estimate of $G_j(t)$ is $\tilde{G}_j(t) = \exp[-\tilde{H}_j(t)]$, where

$$\tilde{H}_j(t) = \sum_{\{i:t(i)\le t\}} \frac{d_{ji}}{n_i} \qquad (3.1.12)$$

is an empirical hazard function.

The marginal survivor function $S(t)$ can be estimated by the product-limit estimate, ignoring cause of death. This is

$$\hat{S}(t) = \prod_{\{i:t(i)\le t\}} \frac{n_i - d_i}{n_i} \qquad (3.1.13)$$

where d_i is the number of deaths at t_i. It is easily seen that $\hat{S}(t) = \hat{G}_1(t) \cdots \hat{G}_J(t)$ (4), provided that there are no tied failure times involving different causes of death.

Alternatively, $H(t)=-\log S(t)$ can be estimated by the empirical hazard function.

Since cause-specific failure probabilities are defined by

$F_j(t)=\int_0^t h_j(u)S(u)du$, $j = 1,\ldots,m$, where

$S(u)=\exp\left\{-\int_0^u\left(\sum_{j=1}^{J}h_j(x)\right)dx\right\}$, the maximum likelihood estimate of $F_j(t)$ is

$$\hat{F}_j(t)=\sum_{\{i:t(i)\le t\}}\frac{d_{ji}}{n_i}\hat{S}(t_{(i-1)}) \tag{3.1.14}$$

where

$\hat{S}(t_{(i-1)})=\prod_{k=1}^{i-1}\left(1-\frac{d_k}{n_k}\right)$, d_k = all deaths at time k (34).

Using a Taylor series linear approximation, it can be shown that for $t_{(k)}\le t<t_{(k+1)}$, the estimated variance of $\hat{F}_j(t)$ (i.e., $\hat{\mathrm{var}}(\hat{F}_j(t))$) equals (12,53)

$$\sum_{i=1}^{k}\hat{\mathrm{var}}(\hat{h}_{ji}\hat{S}(t_{(i-1)}))+2\sum_{i=1}^{k-1}\sum_{i'=i+1}^{k}\hat{\mathrm{cov}}(\hat{h}_{ji}\hat{S}(t_{(i-1)}),(\hat{h}_{ji'}\hat{S}(t_{(i'-1)})), \tag{3.1.15}$$

where

$$\hat{\mathrm{var}}(\hat{h}_{ji}\hat{S}(t_{(i-1)}))=(\hat{h}_{ji}\hat{S}(t_{(i-1)}))^2\left(\frac{(n_i-d_{ji})}{d_{ji}n_i}+\sum_{l=1}^{i-1}\frac{d_l}{n_l(n_l-d_l)}\right)$$

and

$$\hat{\mathrm{cov}}(\hat{h}_{ji}\hat{S}(t_{(i-1)}),\hat{h}_{ji'}\hat{S}(t_{(i'-1)})) = \hat{h}_{ji}\hat{S}(t_{(i-1)})\hat{h}_{ji'}\hat{S}(t_{(i'-1)})\left(-\frac{1}{n_i}+\sum_{l=1}^{i-1}\frac{d_l}{n_l(n_l-d_l)}\right)$$

for $i<i'$. If there is only one cause of failure present, then $1-\hat{G}_1(t)=\hat{F}_1(t)$ and its standard error (SE) ($\hat{F}_1(t)$) would reduce to Greenwood's formula.

3.1.3 Net Probability vs. Crude Probability

The approach 1-G_j accounts for only the probability of failure from cause j and assumes zero for the probability of failure from other causes (12) (net probability). To be at risk of failure at a particular time u, one must survive all causes of failure up to time u to subsequently fail of cause j at time u, thus the term $h_j(u)S(u)$ appears in (3.1.6) (12) (crude probability).

1-$G_j(t)$, compared to $F_j(t)$ of equation (3.1.6), accounts for only the probability of surviving from cause of interest and assumes that of other causes is 1. This leads to an inflated estimate of the proportion of patients who are at risk of failure at time t, which causes 1-$G_j(t)$ to be larger than $F_j(t)$ (12).

If i represents the order of distinct event for cause of interest, $1-\hat{G}_1(t)$ can be expressed for every event,

$$i=1;\quad 1-\left(1-\frac{d_{11}}{n_1}\right)$$

$$i=2;\quad 1-\left(1-\frac{d_{11}}{n_1}\right)\left(1-\frac{d_{12}}{n_2}\right)$$

$$i=3;\quad 1-\left(1-\frac{d_{11}}{n_1}\right)\left(1-\frac{d_{12}}{n_2}\right)\left(1-\frac{d_{13}}{n_3}\right)$$

....

If there is no tied failure times involving different causes of death, and it is assumed that $n_{i'}$ is number at risk when cause 2 $d_{2i'}$ occurs between t_i and t_{i+1}, $\hat{F}_1(t)$ can be expressed as

$$i=1;\quad \hat{F}_1(t_1)=\frac{d_{11}}{n_1}$$

$$i=2;\quad \hat{F}_1(t_2)=\hat{F}_1(t_1)+\frac{d_{12}}{n_2}\left(1-\frac{d_{11}}{n_1}\right)\left(1-\frac{d_{21'}}{n_{1'}}\right)$$

$$i=3;\quad \hat{F}_1(t_3)=\hat{F}_1(t_2)+\frac{d_{13}}{n_3}\left(1-\frac{d_{11}}{n_1}\right)\left(1-\frac{d_{21'}}{n_{1'}}\right)\left(1-\frac{d_{12}}{n_2}\right)\left(1-\frac{d_{22'}}{n_{2'}}\right)$$

....

The difference between $1-\hat{G}_1(t)$ and $\hat{F}_1(t)$ depends on the quantity $\left(1-\frac{d_{2i'}}{n_{i'}}\right)$ prior to time t. If there is no cause 2 in the data, $d_{2i'}$ becomes 0, and $\left(1-\frac{d_{2i'}}{n_{i'}}\right)$ is 1. In that case $\hat{F}_1(t)$ becomes $1-\hat{G}_1(t)$. As the number of deaths by cause 2 $\sum_{\{i':t(i')<t\}} d_{2i'}$ increases, $\prod_{\{i':t(i')<t\}}\left(1-\frac{d_{2i'}}{n_{i'}}\right)$ will decrease from 1. Multiplication by a probability not close to 1 results in an increased difference between $1-\hat{G}_1(t)$ (net probability) and $\hat{F}_1(t)$ (crude probability). If the number of uncensored data increases, it results in an increase in number of deaths by cause 2. It also results in an increased difference between the net probability and the crude probability. Examples of this will be provided in section 4.1 in Chapter 4.

3.2 Heterogeneous Subjects (effects of covariate)

3.2.1 Hazard Functions

Suppose that failure time is continuous and $z(t)$ denotes the value of the regression vector of covariates at time t.

The overall failure rate or hazard function with covariates is defined by

$$h(t;z) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t, z)}{\Delta t}, \quad z = \{z(u); u \ge 0\} .$$

Similarly cause-specific hazard functions are defined by

$$h_j(t;z) = \lim_{\Delta t \to 0} \frac{\Pr[t \le T < t + \Delta t, cause = j \mid T \ge t, z(t)]}{\Delta t}, \tag{3.2.1}$$

for $j = 1, \ldots, m$ (1,2). In words, $h_j(t;z)$ is the instantaneous rate for failure of type j at time t given $\boldsymbol{z}$ and in the presence of the other failure types. Assuming that failure type j must be a unique element of $\{1, 2, \ldots, m\}$ gives

$$h(t;z) = \sum_{j=1}^{m} h_j(t;z) . \tag{3.2.2}$$

With the covariate types under consideration the survivor function,

$$S(t;z) = \exp\left[-\int_0^t h(u;z)du\right] . \tag{3.2.3}$$

The failure time (sub)density function for failure type j, given $\boldsymbol{z}$, is

$$f_j(t;z) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, cause = j | z)}{\Delta t}$$

$$= h_j(t;z) S(t;z), \qquad j = 1, \ldots, m. \tag{3.2.4}$$

Equations (3.2.2) to (3.2.4) show that the likelihood function can be written entirely in terms of the cause-specific hazard functions. Further, these functions are identifiable; that is, they can be estimated from data of the type $(t,j;z)$ without further assumptions.

The likelihood function under an independent censoring mechanism is

$$\prod_{i=1}^{n}\{[h_{ji}(t_i;z_i)]^{\delta_i} S(t_i;z_i)\} = \prod_{i=1}^{n}\left\{[h_{ji}(t_i;z_i)]^{\delta_i} \prod_{j=1}^{m} \exp[-\int_0^{t_i} h_j[u;z(u)]du]\right\}, \quad (3.2.5)$$

where δ_i is a censoring indicator, taking a value of one if failure occurs and a value of zero otherwise (1,2). The likelihood function is completely specified by the cause-specific hazard functions $h_j(t;z)$, $j = 1,\ldots,m$, and can be rearranged into components for each j. In fact, the likelihood factor for $h_j(t;z)$ is precisely the same as would be obtained by treating extraneous causes as censored at their time of failure. This implies that the proportional hazard model of Cox (13,54) can be utilized to model effects of regression variables on cause-specific hazard functions

by treating extraneous causes as censored. It is convenient to denote

$$G_j(t;z) = \exp\left[-\int_0^t h_j(u;z)du\right], \quad j=1,\ldots,m \tag{3.2.6}$$

although these function will not, in general, have any survivor function interpretation for $m>1$.

Proportional hazards modeling in which the cause-specific hazard function at time t depends on $\boldsymbol{z}$ only in terms of the concurrent value $\boldsymbol{z}(t)$ gives

$$h_j(t;z)=h_{0j}(t)\exp(z\beta_j), \quad j = 1,\ldots,m \tag{3.2.7}$$

where $h_{0j}(\cdot)\geq 0$ is arbitrary and β_j, $j = 1,\ldots,m$ are column vectors of cause-specific regression coefficients to be estimated from the data. It is noted that "shape" functions h_{0j} and the regression coefficients β_j have been permitted to vary arbitrarily over the m failure types (1,2).

Let $t_{j1}<\ldots<t_{jdj}$ denote the times of the d_j failures of type j, $j=1,\ldots,m$ and let z_{ji} be the covariate vector for the individual that fails at t_{ji}. The method of partial likelihood then gives

$$L(\beta_1,\ldots,\beta_m) = \prod_{j=1}^{m}\prod_{i=1}^{d_j}\left(\frac{\exp[z_{ji}(t_{ji})\beta_j]}{\sum_{l\in R(t_{ji})}\exp[z_l(t_{ji})\beta_j]}\right) \tag{3.2.8}$$

where $R(t_{ji})$ is the set of study subjects known to be at risk just prior to t_{ji}. Estimation and comparison of the β_j's can be conducted by applying standard asymptotic likelihood techniques individually to the m factors in (3.2.8).

3.2.2 Cause-Specific Failure Probability

The functions $G_j(t,z)$ in (3.2.6) can be estimated at specified $\boldsymbol{z}$.(14,55)

$$\hat{G}_j(t;z) = \exp\left[-\left\{\exp(z\hat{\beta}_j)\sum_{\{i:x(i)\le t\}}[d_{ji}/\sum_{l\in R(t_i)}\exp(z_l\hat{\beta}_j)]\right\}\right], \quad j=1,\ldots,m, \tag{3.2.9}$$

where d_{ji} indicates number of individuals dying at t_{ji}, $R(t_i)$ is the set of indices for the n_i individuals at risk at t_i, and $\hat{\beta}_j$ is the parameter estimate of proportional hazard model for cause j. $1-\hat{G}_j(t,z)$ is the estimate of cause-specific failure probability with covariate effects based on net probability.

Functions of cause-specific failure probabilities based on crude probability, given covariates, is written as (2)

$$F_j(t;z) = \int_0^t h_j(u;z)S(u;z)du \,. \qquad \textbf{(Model 1)}$$

A reasonable estimate of $F_j(t;z)$ is

$$\hat{F}_j(t;z) = \sum_{\{i:t(i)\le t\}} \hat{h}_j(t_{(i)};z)\hat{S}(t_{(i-1)};z)$$

where $\hat{S}(t_{(i-1)};z) = \exp\left[-\left\{\exp(z\hat{\beta})\sum_{k=1}^{i-1}[d_k / \sum_{l\in R(t_k)} \exp(z_l\hat{\beta})]\right\}\right]$ and

$\hat{h}_j(t_{(i)};z) = \dfrac{\exp(z\hat{\beta}_j)}{\sum_{l\in R(t_i)} \exp(z_l\hat{\beta}_j)}$ when a failure from the cause of interest occurs at time t_i, and $\hat{h}_j(t_{(i)};z) = 0$ otherwise. Here, $\hat{\beta}$ is the parameter estimate of proportional hazard model for overall deaths.

Other forms of cause-specific failure probabilities can be derived from the formulation of the absolute cause-specific risk with covariates under the proportional hazards model by Benichou and Gail (15).

Under the assumption that the hazard of extraneous causes, $h_2(t)$, does not depend on $\boldsymbol{z}$, the failure probability for the cause of interest is formulated as

$$F_1(t;z) = \int_0^t h_1(u;z)G_1(u;z)G_2(u)du \ . \qquad \textbf{(Model 2)}$$

A reasonable estimate of this is

$$\hat{F}_1(t;z) = \sum_{\{i:t(i)\le t\}} \hat{G}_{01}(t)^{\exp(z\hat{\beta}_1)} \hat{G}_2(t)\exp(z\hat{\beta}_1)\, d\hat{\Lambda}_{01}(t)$$

where $\hat{G}_2(t)$ is the Kaplan-Meier estimate for extraneous causes, and where $d\hat{\Lambda}_{01}(t) = \dfrac{1}{\sum_{l \in R(t_i)} \exp(z_l\hat{\beta}_1)}$ if a failure of cause of interest occurs at t_i, and $d\hat{\Lambda}_{01}(t) = 0$ otherwise. The summation in model 2 is over individuals at risk at time t_i. $\hat{G}_{01}(t)$ is the product of the terms $\{1 - d\hat{\Lambda}_{01}(t)\}$ over distinct failure times prior to t_i and $\hat{\beta}_1$ is the usual partial likelihood estimator for β_1 (9).

They also mentioned, there is no essential difficulty in allowing $h_2(t;z) = h_{02}(t)\exp(z\beta_2)$, where $h_{02}(t)$ is an arbitrary unspecified base-line hazard function for cause 2 at time t, and β_2 is provided functionally independent of β_1. In that case, $F_1(t;z)$ can be written as

$$F_1(t;z) = \int_0^t h_1(u,z)G_1(u,z)G_2(u,z)du \,. \qquad \textbf{(Model 3)}$$

A reasonable estimate of this is

$$\hat{F}_1(t;z) = \sum_{\{t:t(l)\le t\}} \hat{G}_{01}(t)^{\exp(z\hat{\beta}_1)}\, \hat{G}_{02}(t)^{\exp(z\hat{\beta}_2)} \exp(z\hat{\beta}_1)\, d\hat{\Lambda}_{01}(t)$$

where $\hat{G}_{02}(t)$ and $\hat{\beta}_2$ are estimated in the same way as $\hat{G}_{01}(t)$ and $\hat{\beta}_1$ in model 2.

$\hat{G}_j(t;z)$ in Eq (3.2.9) and $\hat{G}_{0j}(t)^{\exp(z\hat{\beta}_j)}$ in models 2 and 3 are same formulas if there is no ties. The former is derived by the empirical method, and the latter is done by the Kaplan-Meier method. Notice that the estimate of hazard function in each model 1, 2, and 3, $\dfrac{\exp(z\hat{\beta}_j)}{\sum_{l\in R(t_i)} \exp(z_l\hat{\beta}_j)}$, is identical. Therefore, the estimates of CSFP at time t by the three models vary depending on the composition of the overall survival prior to time t, mainly how survival of cause 2 influences the overall survival prior to time t. Model 1 and model 3 would provide the same estimates, given

no ties with different failure types and valid proportionality assumptions for cause 1 and cause 2.

Equations in model 2 and model 3 are generalizations of equation (3.1.8). Model 3 is more practical than model 2, because $\hat{F}_1(t,z)$ is affected by the survival functions $\hat{G}_1(t)$ and $\hat{G}_2(t)$, and both survival functions are influenced by covariates.

3.2.3 Checking for Appropriateness of β Estimators

A specialization of (3.2.7) that is expected to give rise to more valid β estimators, if applicable, is that given by

$$h_j(t;z)=h_{0j}(t)e^{\gamma_j}\exp(z\beta_j), \qquad j = 1,\ldots,m. \qquad (3.2.10)$$

This time, the cause-specific hazards are assumed to be proportional to each other with proportionality factors e^{γ_j} (for uniqueness set γ_1=0) (2). This proportionality assumption can be checked graphically on the basis of plots of $\log[-\log\hat{G}_j(t;z)]$ versus t, for specified $\boldsymbol{z}$, where the $\hat{G}_j$ are estimated from an analysis of model (3.2.7). Such plots

should be separated by approximately a constant difference for various values of j under (3.2.10).

3.2.4 Net Probability vs. Crude Probability

As the same way as in section 3.1.3, the difference between the estimate of net probability and that of crude probability considering covariate effects depends on the quantity $\left(1-\frac{1}{\sum_{l\in R(t_i)}\exp(z_l\beta_2)}\right)^{\exp(\tilde{z}\hat{\beta}_2)}$ prior to time t, where $\hat{\beta}_2$ is the usual partial likelihood estimator for β_2, a coefficient of a covariate for cause 2, and $\tilde{z}$ is a given value to estimate the CSFP. If this quantity is nearer 0, the difference between the two probabilities is greater. Because $0<\left(1-\frac{1}{\sum_{l\in R(t_i)}\exp(z_l\beta_2)}\right)<1$, given competing risks, and $\sum_{l\in R(t_i)}\exp(z_l\beta_2)$ is fixed in the given data, the larger $\tilde{z}\beta_2$ (i.e., the power) is, the greater the difference is. For example, if a covariate is positively related to cause 2, the difference is increased according to the increasing

value of the covariate. If a covariate is negatively related to cause 2, the direction is in the opposite way. When coefficients of a covariate for cause 2 vary among different populations, the stronger the association of a covariate with cause 2, the larger the difference is.

3.3 Study Methods

First, I will compare the estimates of net probability and crude probability for simulated data and for prostate cancer patients in Hawaii assuming homogeneity in characteristics. Then I will investigate how the estimates of the two probabilities differ according to the proportion censored and the proportion of deaths due to the cause of interest in simulated data, and how great the difference is in the data of prostate cancer patients in Hawaii.

Second, I will compare the estimates of net probability and models 1-3 for crude probability for prostate cancer patients of Hawaii considering heterogeneity in characteristics. I will investigate how the estimates of

net probability and each model for crude probability differ depending on values of covariates.

Third, I will check the proportionality assumption and suggest an alternative way of estimating the CSFP if the assumption is violated.

CHAPTER 4

RESULTS

In this chapter, I shall examine the difference between net probability and crude probability in estimating cause-specific failure probability. Section 4.1 uses generated and real data under the assumption of homogeneous subjects in characteristics. Section 4.2 uses real data under the assumption of heterogeneous subjects in characteristics.

4.1 Homogeneous Subjects

4.1.1 Simulated Data

A simulated data set was created by a using SAS program (Appendix I) under the distribution free assumption. The total sample size was 500, and there were 16 years of person years (survival time) where the events were distributed uniformly. These assumptions are similar to the situation in the real data of prostate cancer patients in Hawaii. The binary variable status (censored or uncensored) was pseudorandomly assigned 1 or 0 by the uniform distribution

(*ranuni*). Specifically, assuming the proportion of censored data is 20%,

if *ranuni(seed)*<0.2 then status=1;

if *ranuni(seed)*>=0.2 then status=0, where *seed* is a value which initializes the random number stream.

The cause variable (cause of interest=cause1, other causes=cause2) was generated by assigning 1 or 2 by a random table generator (*rantbl*). For example, when the status was noncensored (status=0),

cause=*rantbl(seed*, 0.8, 0.2).

The cause was assigned missing when the status was censored (status=1). Here I assumed 80% of deaths were due to cause 1. Six data sets were generated. The proportion censored was assigned as 0%, 20% and 50%, and then the proportion of cause of interest was assigned as 80% and 50% for the noncensoring part of each set.

For these sample sets, the net probability and the crude probability were estimated as follows,

$1-\hat{G}_1(t)=1-\prod_{\{i:t(i)\le t\}}(1-\frac{d_{1i}}{n_i})$ (where n_i is the number of individuals at risk just prior to t_i and d_{ji} is the number of individuals

dying from cause j at t_i) for the net probability, $\hat{F}_1(t) = \sum_{\{i:t(i)\le t\}} \frac{d_{1i}}{n_i}\hat{S}(t_{(i-1)})$ (where $\hat{S}(t_{(i-1)}) = \prod_{k=1}^{i-1}\left(1-\frac{d_k}{n_k}\right)$, d_k = all deaths at time k) for the crude probability. Tables 1-3 show the results of $1-\hat{G}_1(t)$, $\hat{F}_1(t)$, and standard errors for four time points representing the quartiles (time 1; 4.1 years, time 2; 7.9 years, time 3; 12.3 years, time 4; 16.0 years) of the above data set, and Table 4 shows the summary of how much the estimates increased when the net probability is used for the cause-specific failure probabilities instead of the crude probability. Figures 1-3 show their graphical presentations. As explained in chapter 3, it is evident that the differences between $1-\hat{G}_1(t)$ and $\hat{F}_1(t)$ increased as the proportions of the censored decreased in the data set. If the proportion censored is the same, the difference increases when the proportion of deaths due to the cause of interest is lower (i.e., the proportion of other causes is higher).

4.1.2 Description of Data

The real data used for this study came from a case-control investigation of prostate cancer carried out on the island of Oahu, Hawaii, during the period 1977-1983 by the Cancer Research Center of Hawaii (56,57). Of the original study, only cases were used for this study. Cases included only histologically confirmed cancer cases and were identified thorough the rapid-reporting system of the population-based Hawaii Tumor Registry, a member of the national Surveillance, Epidemiology, and End Results (SEER) Program. For practical reasons, the rapid-reporting system covers only the major medical facilities on Oahu, but these hospitals included over 87 per cent of all newly diagnosed cases of prostatic cancer with five major ethnic groups (Caucasian, Japanese, Chinese, Filipino, Hawaiian) in the state. Only Oahu residents were eligible for study. Data included dietary, occupational, medical, social, and demographic information. The causes of death were based on deaths identified in the death certificate files through December 1991. Out-migration from Hawaii is limited (58). In a random sample of 2,018 members from a separate Cancer

Research Center cohort of 41,400 composed of a random 2% of households in the state surveyed by the Hawaii Department of Health, most of the out-migration occurred within 10 years of entry and stabilized at 10.7% after the first 13 years. As of January 1992, only 1% of the sample could not be traced. Because the subjects of the present study are prostate cancer patients, the out-migration rate for these people might be less than in the cohort substudy of out-migration.

Among 451 prostate cancer patients, 12 cases whose survival times were zero were excluded, as this study deals with survival analysis. Five subjects whose information on stage of disease was missing were also excluded since this information is a covariate of interest. More detailed description of the data will be given in section 4.2.1.

Thus a total of 434 cases were included in the study, including 273 failures (uncensored cases, 62.9%) and 161 censored cases (37.1%). Among the uncensored cases, the proportion of deaths due to prostate cancer and those due to other causes were 37.0 percent (101 cases) and 63.0 percent (172 cases), respectively.

4.1.3 Analysis of Data under Homogeneity Assumption

For the data of prostate cancer patients in Hawaii, $1-\hat{G}_1(t)$, $\hat{F}_1(t)$, and their standard errors were estimated under the assumption that patients are homogeneous in characteristics. Tied failure times involving different causes of death (thereafter called "tied failure times") are ignored, because there are few in the data and they are not especially influential in estimating the CSFP. When a small amount of time is added to the failure time of death by cause 2 to avoid tied failure times, there is no remarkable difference in results compared to those estimated by ignoring the tied failure times. Table 5 shows the results for four time points of person-years, with the graphical presentation in Figure 4. The estimate of net probability of death from prostate cancer at time t, $1-\hat{G}_1(t)$, is increased that by crude probability, $\hat{F}_1(t)$, by 6.4 percent at 3.71 years of survival time, increasing up to 36.5% at 15.45 years survival time.

4.2 Heterogeneous Subjects (effects of covariate)

4.2.1 Analysis of Covariates in Proportional Hazards Model

Two covariates, age and stage, were selected to be included in the analysis. Stage was categorized into 'localized', 'regional', and 'distant' stages. The stage 'in situ' was included in 'localized', because it included only one case. The reference group was 'localized', and two dummy variables representing 'regional' and 'distant' were included in the analysis of proportional hazards model. Continuous age was also included in Cox's proportional hazards model. The coefficients of the covariates were obtained from the PHREG routine of SAS. β's for the cause of interest were obtained by treating extraneous causes as censored, vice versa. Table 6 gives the details of model fitting to the prostate cancer data. Column 2 shows the effects of the explanatory variables on the overall survival (prostate cancer and other causes) while column 3 and 4 present those on the causes separately.

Age is a significant variable for overall and cause-specific hazards with older age being associated with higher

risk. The effect is lower for the risk of prostate cancer death than for the risk of death by "other causes". "Other causes" contains deaths from cardiovascular disease, respiratory disease, and other specified and unspecified causes. Stage had a significant effect with advanced stage of cancer giving poorer survival. However, for the risk of death by "other causes", the 'regional' stage did not have a significant effect compared to the 'localized' stage. It means that the effect of the stage 'regional' on the hazard of other causes is not significant compared to that of the stage 'localized' on the hazard. Overall, stage affected the risk of death by prostate cancer more than age did, while age affected death due to "other causes" more than stage did. In other words, if a patient is older, he has higher risk to die with other causes rather than with prostate cancer, and if a patient is in advanced stage of prostate cancer, e.g., 'distant', he tends to die with prostate cancer rather than with other causes. The effect of each variable on the risk of overall death is intermediate between the two cause-specific hazards.

To explain cause-specific failure probability with covariate effects, these coefficients will be used in Section 4.2.2. Graphical checking for the proportionality assumption will be made in Section 4.2.3, and also an alternative way to measure CSFP will be suggested when the assumption is violated.

4.2.2 Analysis of Cause-Specific Failure Probability

The net probability and the crude probability (models 1-3) due to prostate cancer (cause of interest, cause 1) were estimated and presented in Tables 7-10 and in Figures 5-13. The net probability was estimated by

$$1-\hat{G}_j(t;z)=1-\exp\left[-\left\{\exp(z\hat{\beta}_j)\sum_{\{i:t(i)\le t\}}[d_{ji}/\sum_{l\in R(t_i)}\exp(z_l\hat{\beta}_j)]\right\}\right]$$ (section 3.2.2), and models 1-3 for the crude probability were estimated by

$$\hat{F}_j(t;z)=\sum_{\{i:t(i)\le t\}}\hat{h}_j(t_{(i)};z)\hat{S}(t_{(i-1)};z)$$ for model 1 (section 3.2.2),

$$\hat{F}_1(t;z)=\sum_{\{i:t(i)\le t\}}\hat{G}_{01}(t)^{\exp(z\hat{\beta}_1)}\hat{G}_2(t)\exp(z\hat{\beta}_1)\,d\hat{\Lambda}_{01}(t)$$ for model 2 (section 3.2.2),

$$\hat{F}_1(t;z)=\sum_{\{i:t(i)\le t\}}\hat{G}_{01}(t)^{\exp(z\hat{\beta}_1)}\hat{G}_{02}(t)^{\exp(z\hat{\beta}_2)}\exp(z\hat{\beta}_1)\,d\hat{\Lambda}_{01}(t)$$ for model 3 (section 3.2.2).

The given values of covariates for cause-specific failure probabilities are 66, 71, 77, for ages, and 'localized', 'regional', 'distant' for stages. The three ages represent the approximate 25th, 50th, 75th percentiles of the distribution of all cases. Two dummy variables are used for 'regional' and 'distant' for stages as in the analysis of covariates with the proportional hazards model in Section 4.2.1.

The estimates of net probability were always higher than those of crude probability, models 1 to 3. To measure the difference of the two estimates, the ratio, $\frac{\textit{Estimate of net probability}}{\textit{Estimate of crude probabilities from Model 1,2, or 3}}$, was calculated. The estimate ratios differ depending on the values of the covariates and the model used for the crude probability as shown Table 11. This estimate ratio is useful in explaining the characters of each model for crude probability.

As mentioned in Chapter 3, the estimate of failure probability by cause 1 in each model varies depending on the overall survival, mainly how survival of cause 2 influences the overall survival.

The ratios of the estimate of net probability to that of model 3 for crude probability were examined for each value of covariates in Table 11. I will first discuss these because, compared to the other two models, model 3 shows the influence of covariates on survival from cause 2 which comprises overall survival with cause 1. We can see the ratio is increased if age is increased for every stage. However, if we compare the ratio of 'localized' and 'regional' in same ages, the estimate ratio of 'regional' is lower than that of 'localized'. The reason is that since the 'localized' and 'regional' stages are regarded as same in the survival for cause 2 (The coefficient of 'regional' is not significant in Cox's model for cause 2, see Table 6.), it results in a decreased estimate ratio in stage 'regional'. In other words, since the risk of 'regional' is not explanatory for the hazard of cause 2 in this model, even though the hazard of cause 1 increases (from stage

'localized' to stage 'regional'), it resulted in making the failure probability for cause 1 to increase. Therefore, the estimate ratio is decreased in the 'regional' stage.

The failure probability of cause 1 by model 2 shows an interesting trend. The estimate ratio is slightly decreasing when age increases and stage moves from 'localized' to 'regional', and 'distant'. It is noticed that the estimates by model 2 are quite different from those by model 1 or model 3 in ages 66 and 77 (Figures 5-7, 11-13), and close to them in age 71, the average age (Figures 8-10). The reason is that the effect of fixed hazard of cause 2 affects the failure probability of cause 1 relatively more or less according to actual hazard of cause 2 being decreased or increased from the average age, respectively. While in model 3 the estimate ratio is the smallest in 'regional' among the three stages within same age, in model 2, the estimate ratio of 'regional' was intermediate because any effect of covariate is not accounted for cause 2.

In model 1 survival from cause 1 and that of cause 2 are mixed in overall survival. In this case, the hazard of

cause 1 acts in two ways, one for the net hazard of cause 1, another for the hazard of cause 1 being mixed into overall survival. Therefore, the estimate ratio is influenced not only by the hazard of cause 2 but also from the effect how the hazard of cause 1 influenced the overall hazard. Overall, CSFP by model 1 and model 3 are close as seen in graphic presentations (Figures 5-13). The estimate ratio is increased as age is increased. As stage moves from 'localized' to 'regional', the estimate ratio is decreased in age 66 and 71, but it shows irregular trends depending on time in age 77. From the 'regional' to 'distant' stages, the estimate ratio decreased in age 66, increased in age 71, and 77 under model 1. Actually, if there were no ties, and the covariates fit the proportionality assumption, the estimates of model 1 and model 3 would have the almost same values. The difference in the results of the two models strongly suggests violation of the proportionality assumption. This will be discussed in the next section.

4.2.3 Checking of the Proportionality Assumption for Hazards

Figures 14-22 present graphical examinations of proportionality assumption of hazards for cause 1 and cause 2 by age and stage. It is easily seen that the proportionality assumption does not hold for these data, because the slopes for cause 1 and cause 2 are not separated by a constant difference for various given values of covariates. Therefore, stratification by stage was made and a proportional hazard model was applied for cause 1 and cause 2 for each stage. Table 12 gives parameter estimates of age for overall (prostate cancer and other causes) and cause-specific hazards by stage and appropriate methods to estimate crude probability of death from prostate cancer. Actually age is not significant in the proportional hazards model for cause 1 in stages 'regional' and 'distant'. An alternative method is required to estimate cause-specific failure probability. A reasonable way is to apply the method of estimation of CSFP differently according to stages. For the 'localized' stage, model 1 and model 3 may be applied, since age explains both hazards of cause 1 and cause 2. For stages 'regional' and 'distant', a

modification to model 2 can be applied. Since age does not fit in the proportional hazard model for cause 1, the hazard of cause 1 can be regarded as fixed, which can be estimated just by non-parametric methods. The estimate of failure probability by cause 1 in stage 'regional' or 'distant' can be expressed as

$$\hat{F}_1(t,z) = \sum_{\{i:t(i)\le t\}} \hat{h}_1(t)\hat{G}_1(t)\hat{G}_{02}(t)^{\exp(z\hat{\beta}_2)} ,$$

where $\hat{h}_1(t) = \frac{d_{1i}}{n_i}$ (see section 3.1.2), and $\hat{G}_1(t)$ is the Kaplan-Meier estimate for cause 1 and $\hat{G}_{02}(t)^{\exp(z\hat{\beta}_2)} = \prod_{\{i:t(i)<t\}} \left(1 - \frac{1}{\sum_{l\in R(t_i)} z_l\hat{\beta}_2}\right)^{\exp(z\hat{\beta}_2)}$. When covariate effects are ignored ($\hat{\beta}_1 = 0$), $\hat{G}_{01}(t)^{\exp(z\hat{\beta}_1)} = \prod_{\{i:t(i)<t\}} \left(1 - \frac{1}{\sum_{l\in R(t_i)} z_l\hat{\beta}_1}\right)^{\exp(z\hat{\beta}_1)}$ reduces to the Kaplan-Meier estimate.

Checking the proportionality assumption for age in cause 2 were performed graphically for stages 'regional' and 'distant'. They showed that the proportionality assumption is valid.

In graphical examination for the proportionality assumption for cause 1 and cause 2 in stage 'localized' and age 71, there is a crossing in the survival curves around 1 year (Figure 23). Therefore, two time zones were made, one for less than 1.2 year (time zone one), one for greater than or equal to 1.2 year (time zone two). Now slopes by cause 1 and cause 2 do not cross for a given age in time zone two (Figure 24).

To get the cause-specific failure probability for cause 1, $\hat{F}_j(t) = \sum_{\{i:t(i)\le t\}} \frac{d_{ji}}{n_i}\hat{S}(t_{(i-1)})$ (eq. 3.1.14) is applied for the time zone one and model 1 and model 3 are applied for time zone two. For the analysis of time zone one, survival times equal to or greater than 1.2 years are treated as censored. Survival times less than 1.2 years are excluded in the analysis of time zone two. However, the probability of survival prior to time of 1.2 years (0.98616) is multiplied by the estimated probability of survival for time zone two. CSFP in time zone one are added those in time zone two for the final distribution of cause-specific failure probability (tables 13-14).

Table 13 presents estimates of failure probability due to cause 1 by this alternative method. It is noticed that the estimates by model 1 and model 3 are almost the same in time zone two of 'localized'. It seems that the proportionality assumption of Cox's model is appropriate for this stratum.

Estimates of failure probability due to other causes (cause 2) by the alternative method is presented in Table 14. There is a little difference between the estimate by model 1 and that by model 3 in time zone two of the 'localized' stage. This suggests that the proportionality assumption does not perfectly fit in this stratum. However, it is acceptable to apply the proportional hazards model in this stratum, because the difference is negligible.

The comparison of estimate ratios of failure probabilities due to cause 1 and cause 2 from net probability and model 3 of crude probability was performed in time zone two of stage 'localized' (Table 15). It shows that the difference increases according to increasing value of the covariate (age) when the covariate is positively related to cause 2. It is also noticed that the more the

association of a covariate is related to extraneous causes the larger the difference is.

Table 1

Estimates and standard errors for net and crude probabilities by proportion of deaths due to the cause of interest. Simulated data, under the assumption of no censoring.

time* (person-years)	proportion of cause for interest							
	80%				50%			
	$1-\hat{G}_1(t)$	SE	$\hat{F}_1(t)$	SE	$1-\hat{G}_1(t)$	SE	$\hat{F}_1(t)$	SE
4.1	.200	.018	.194	.018	.116	.015	.108	.014
7.9	.423	.023	.394	.022	.293	.023	.244	.020
12.3	.677	.023	.602	.022	.519	.028	.378	.022
16.0	.993	.007	.802	.018	.944	.031	.500	.022

* Quartiles of survival time.

Table 2

Estimates and standard errors for net and crude probabilities by proportion of deaths due to the cause of interest. Simulated data, under the assumption of 20% censored.

time* (person-years)	proportion for cause of interest							
	80%				50%			
	$1-\hat{G}_1(t)$	SE	$\hat{F}_1(t)$	SE	$1-\hat{G}_1(t)$	SE	$\hat{F}_1(t)$	SE
4.1	.167	.017	.164	.017	.104	.014	.098	.013
7.9	.357	.023	.338	.022	.241	.022	.208	.019
12.3	.577	.025	.529	.022	.424	.028	.332	.022
16.0	.980	.013	.799	**	.914	.045	.499	**

* Quartiles of survival time.

** Not estimable.

Table 3

Estimates and standard errors of net and crude probabilities by proportion of deaths due to the cause of interest. Simulated data, under the assumption of 50% censored.

time* (person-years)	proportion for cause of interest							
	80%				50%			
	$1-\hat{G}_1(t)$	SE	$\hat{F}_1(t)$	SE	$1-\hat{G}_1(t)$	SE	$\hat{F}_1(t)$	SE
4.1	.111	.015	.110	.014	.077	.013	.075	.012
7.9	.259	.022	.254	.021	.167	.019	.157	.017
12.3	.418	.027	.399	.023	.280	.026	.243	.022
16.0	.786	.046	.689	**	.644	.069	.438	**

* Quartiles of survival time.

** Not estimable.

Table 4

Percentage increase in the net probability estimate over that of crude probability due to the cause of interest with simulated data under different proportions censored

time* (person-years)	cause of interest-->	Status: Noncensored 80%	Noncensored 50%	20% censored 80%	20% censored 50%	50% censored 80%	50% censored 50%
4.1		2.9	7.6	2.1	5.8	1.0	2.8
7.9		7.3	20.0	5.6	16.0	2.1	6.8
12.3		12.4	37.3	9.1	27.8	4.9	15.1
16.0		23.8	88.8	22.6	83.1	14.1	47.2

* Quartiles of survival time.

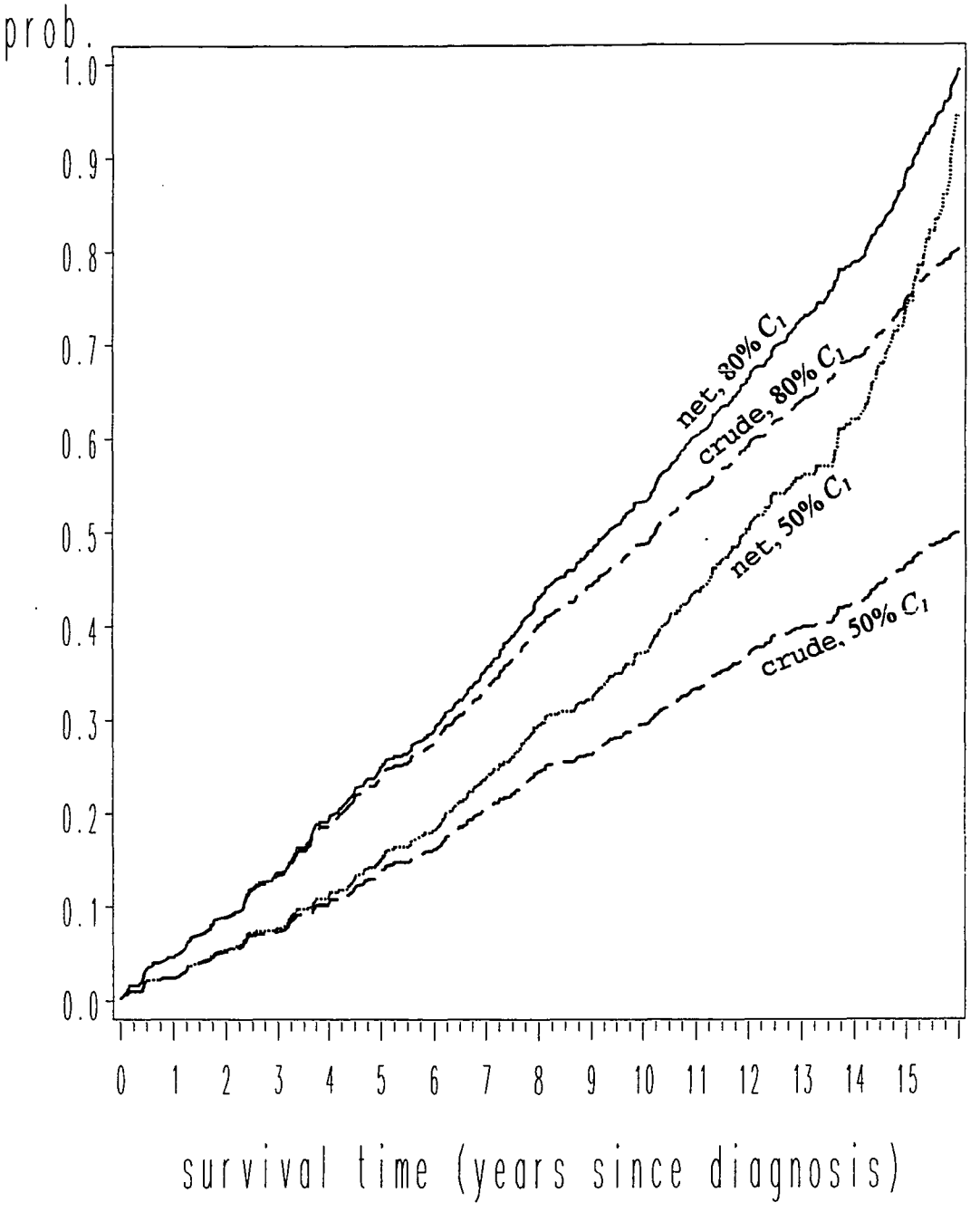


Figure 1. Comparison of estimates of net probability and crude probability. Simulated data, no censoring, proportion of cause of interest (C1); 80% 50%, each.

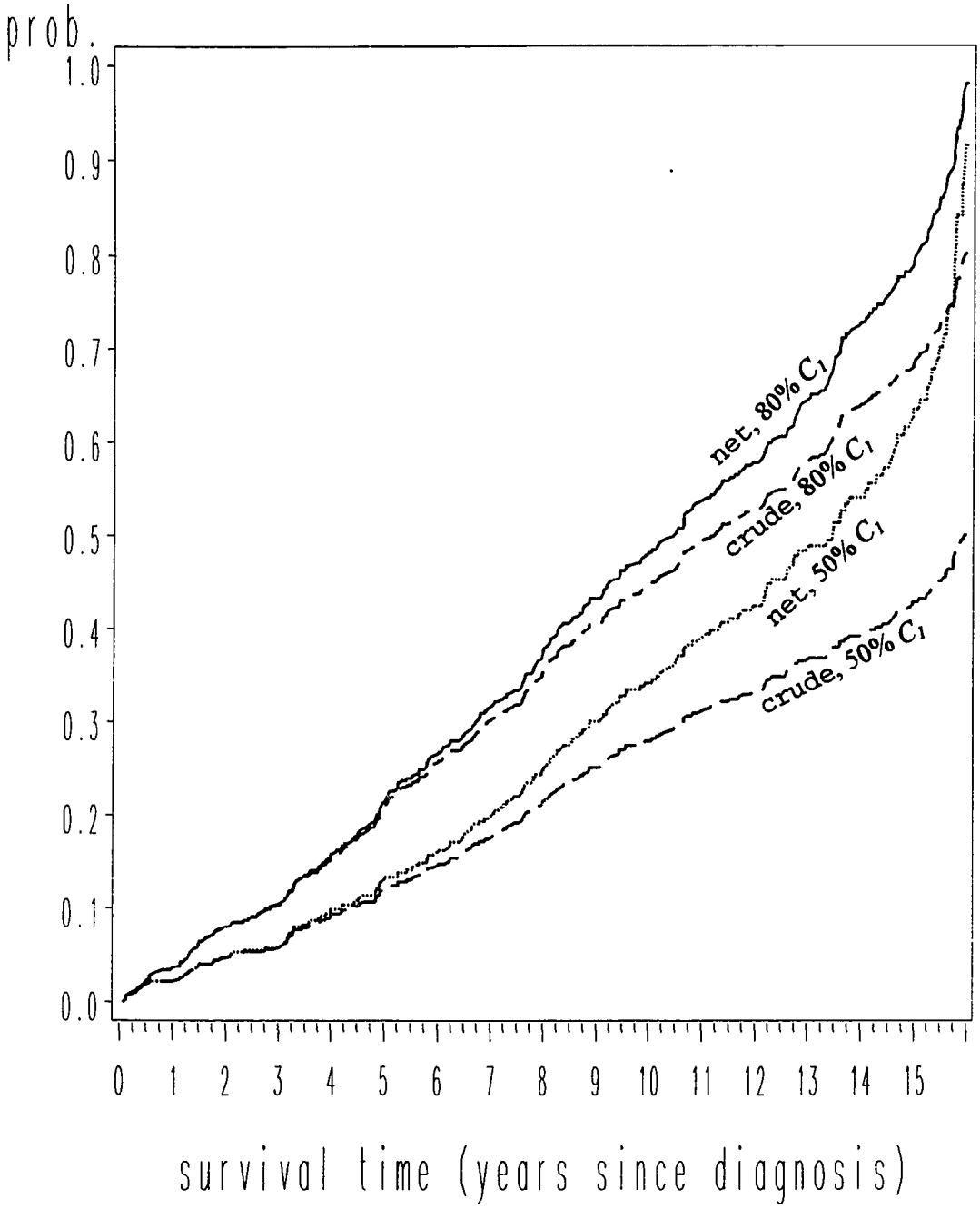


Figure 2. Comparison of estimates of net probability and crude probability. Simulated data, 20% censored, proportion of cause of interest (C1); 80% 50%, each.

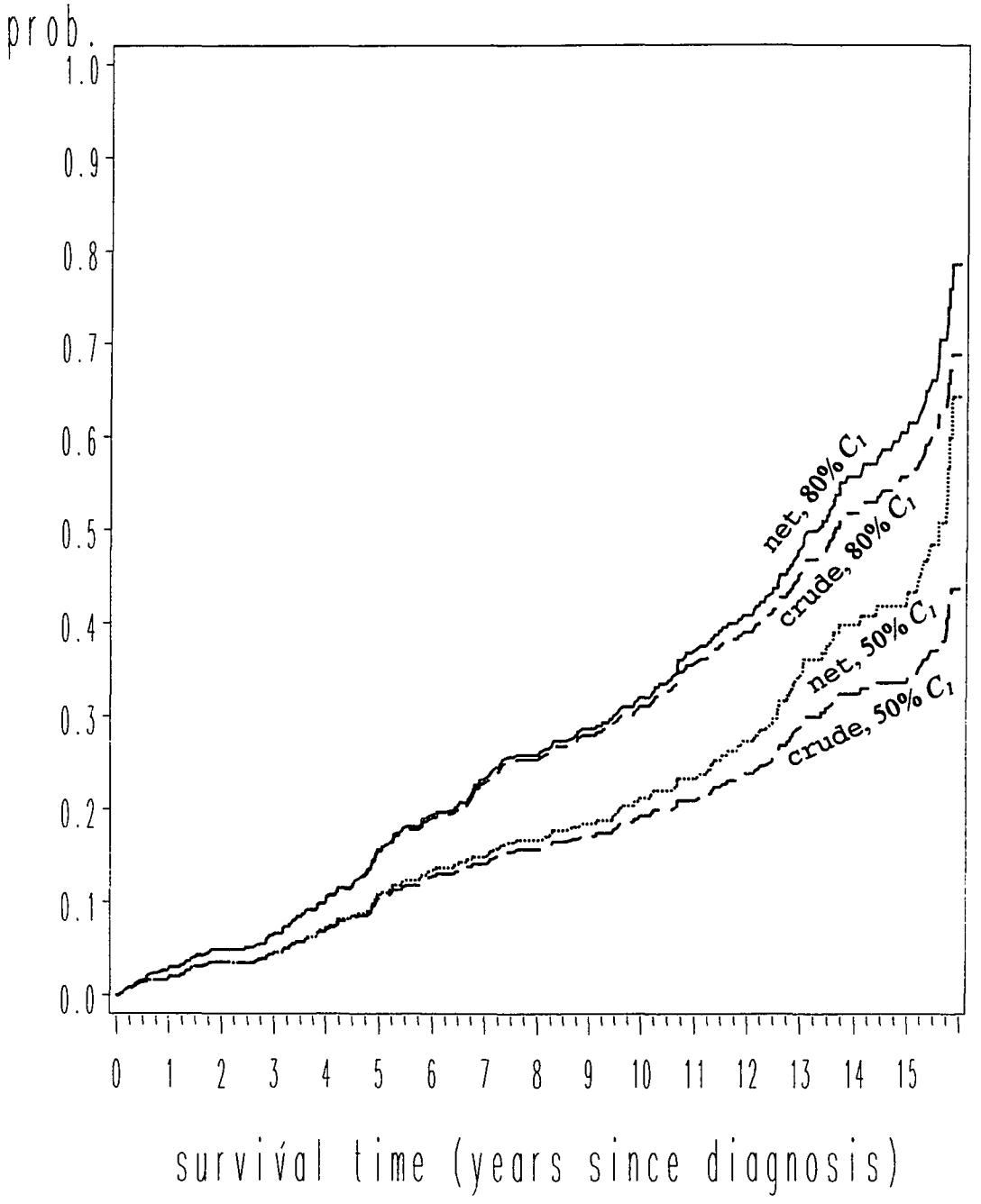


Figure 3. Comparison of estimates of net probability and crude probability. Simulated data, 50% censored, proportion of cause of interest (C1); 80% 50%, each.

Table 5

Estimates of net probability and crude probability of death from prostate cancer for four time points. Prostate cancer patients in Hawaii under the assumption of homogeneity in characteristics.

time*	Methods of Estimation				
(person-	Net probability		Crude probability		
years)	$1-\hat{G}_1(t)$	SE	$\hat{F}_1(t)$	SE	Difference(%)
3.71	.116	.016	.109	.015	6.4
7.66	.233	.023	.198	.019	17.7
11.11	.271	.025	.221	.020	22.6
15.45	.363	.042	.266	.019	36.5

* Quartiles of survival time.

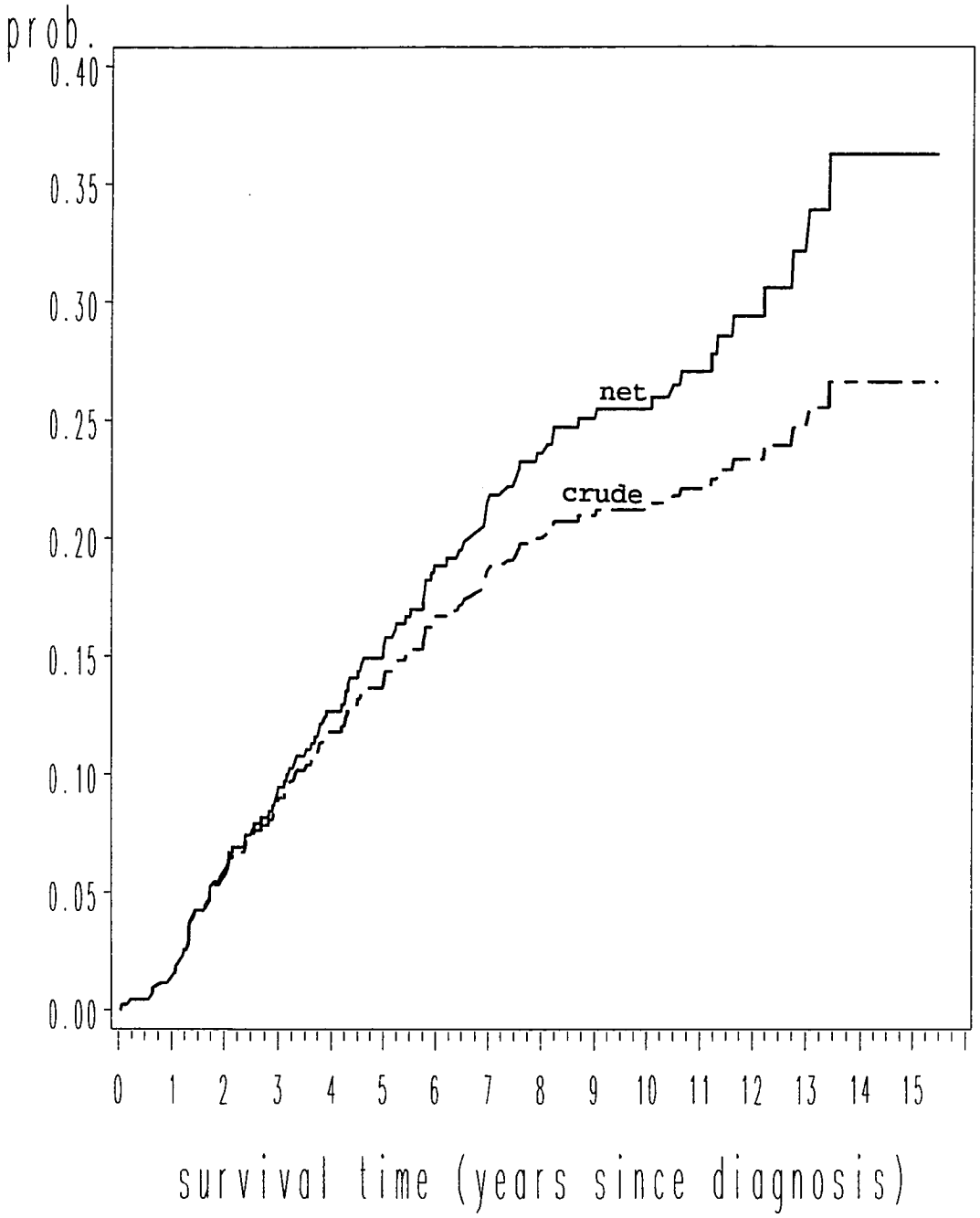


Figure 4. Comparison of estimates of net probability and crude probability. Prostate cancer patients in Hawaii, under the assumption of homogeneous in characteristics.

Table 6

Estimates of coefficients (β,β_j) under proportional hazards models for overall (prostate cancer and other causes) and cause-specific hazards. Standard errors in parentheses.

		Cause-specific	
Covariate	Overall	Prostate Cancer	Other causes
Age	0.072(.009)*	0.031(.014)*	0.097(.011)*
Stage			
Regional vs. localized	0.560(.176)*	1.413(.275)*	0.062(.247)
Distant vs. localized	1.075(.142)*	2.069(.231)*	0.424(.206)*

*p<0.05

Table 7

*Estimates of net probability of death from cause 1. For four time points, by age and stage** of prostate cancer patients.*

Time*	'Localized'	'Regional'	'Distant'
		Age=66	
3.71	.0409	.1577	.2815
7.66	.0936	.3323	.5406
11.11	.1156	.3963	.6217
15.45	.1806	.5589	.7933
		Age=71	
3.71	.0477	.1819	.3206
7.66	.1086	.3765	.5973
11.11	.1338	.4458	.6791
15.45	.2078	.6160	.8417
		Age=77	
3.71	.0572	.2151	.3727
7.66	.1295	.4344	.6663
11.11	.1591	.5094	.7463
15.45	.2450	.6849	.8919

* Quartiles of survival time.

** Age and stage were the covariates in the proportional hazards model for cause 1.

Table 8

*Estimates of model 1 for crude probability of death from cause 1. For four time points, by age and stage** of Prostate cancer patients.*

Time*	'Localized'	'Regional'	'Distant'
		Age=66	
3.71	.0395	.1555	.2809
7.66	.0849	.3142	.5187
11.11	.1015	.3651	.5806
15.45	.1436	.4761	.6875
		Age=71	
3.71	.0451	.1745	.3072
7.66	.0934	.3325	.5209
11.11	.1098	.3768	.5654
15.45	.1480	.4601	.6252
		Age=77	
3.71	.0521	.1959	.3304
7.66	.1016	.3397	.4939
11.11	.1163	.3715	.5165
15.45	.1455	.4172	.5362

* Quartiles of survival time.

** Age and stage were the covariates in the proportional hazards models for overall causes (cause 1 and cause 2) and cause 1.

Table 9

*Estimates of model 2 for crude probability of death from cause 1. For four time points, by age and stage** of Prostate cancer patients.*

Time*	'Localized'	'Regional'	'Distant'
		Age=66	
3.71	.0382	.1478	.2648
7.66	.0783	.2814	.4645
11.11	.0917	.3207	.5146
15.45	.1238	.4028	.6037
		Age=71	
3.71	.0445	.1705	.3020
7.66	.0909	.3196	.5158
11.11	.1063	.3623	.5665
15.45	.1428	.4487	.6517
		Age=77	
3.71	.0535	.2019	.3518
7.66	.1085	.3703	.5795
11.11	.1266	.4165	.6292
15.45	.1691	.5020	.7064

* Quartiles of survival time

** Age and stage were the covariates in the proportional hazards model for cause 1.

Table 10

*Estimates of model 3 for crude probability of death from cause 1. For four time points, by age and stage** of prostate cancer patients.*

Time*	'Localized'	'Regional'	'Distant'
		Age=66	
3.71	.0397	.1534	.2704
7.66	.0859	.3072	.4852
11.11	.1030	.3571	.5408
15.45	.1466	.4687	.6386
		Age=71	
3.71	.0454	.1737	.3001
7.66	.0947	.3322	.5043
11.11	.1114	.3784	.5486
15.45	.1500	.4697	.6126
		Age=77	
3.71	.0525	.1982	.3315
7.66	.1021	.3505	.5033
11.11	.1163	.3867	.5299
15.45	.1430	.4434	.5567

* Quartiles of survival time.

** Age and stage were the covariates in the proportional hazards models for cause 1 and cause 2.

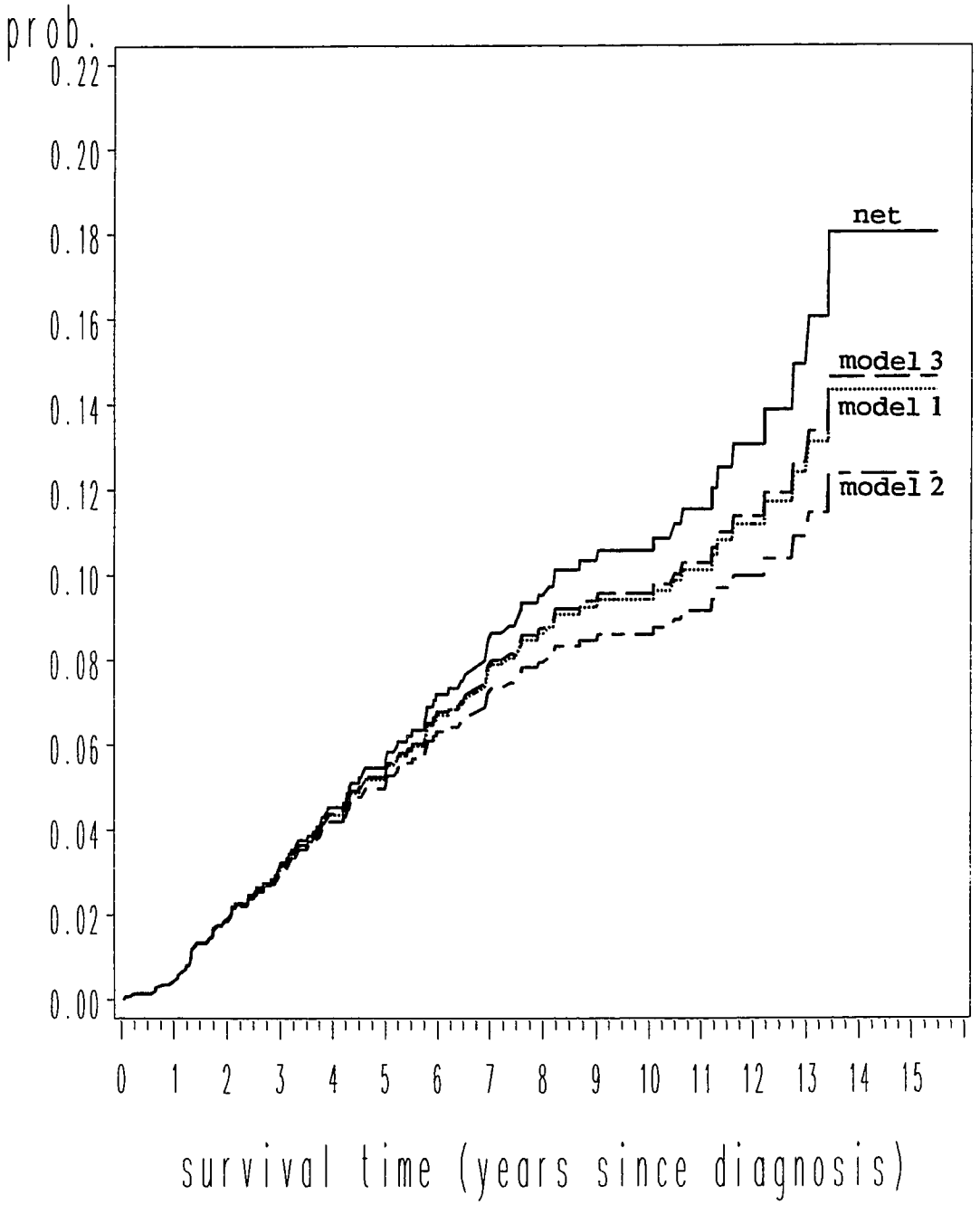


Figure 5. Comparison of estimates of net probability and models 1-3 forcrude probability. Prostate cancer patients in Hawaii, age;66, stage;localized.

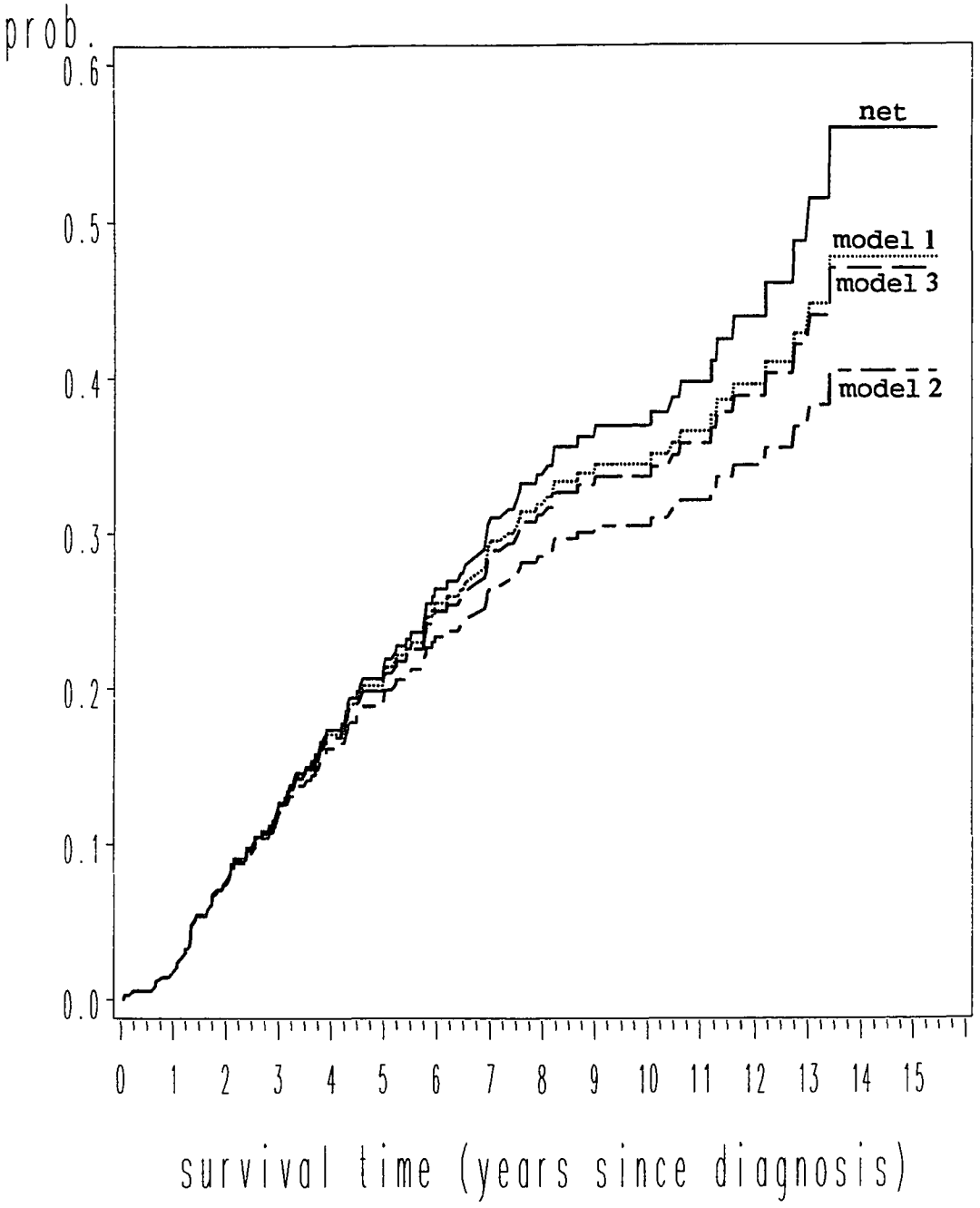


Figure 6. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;66, stage;regional.

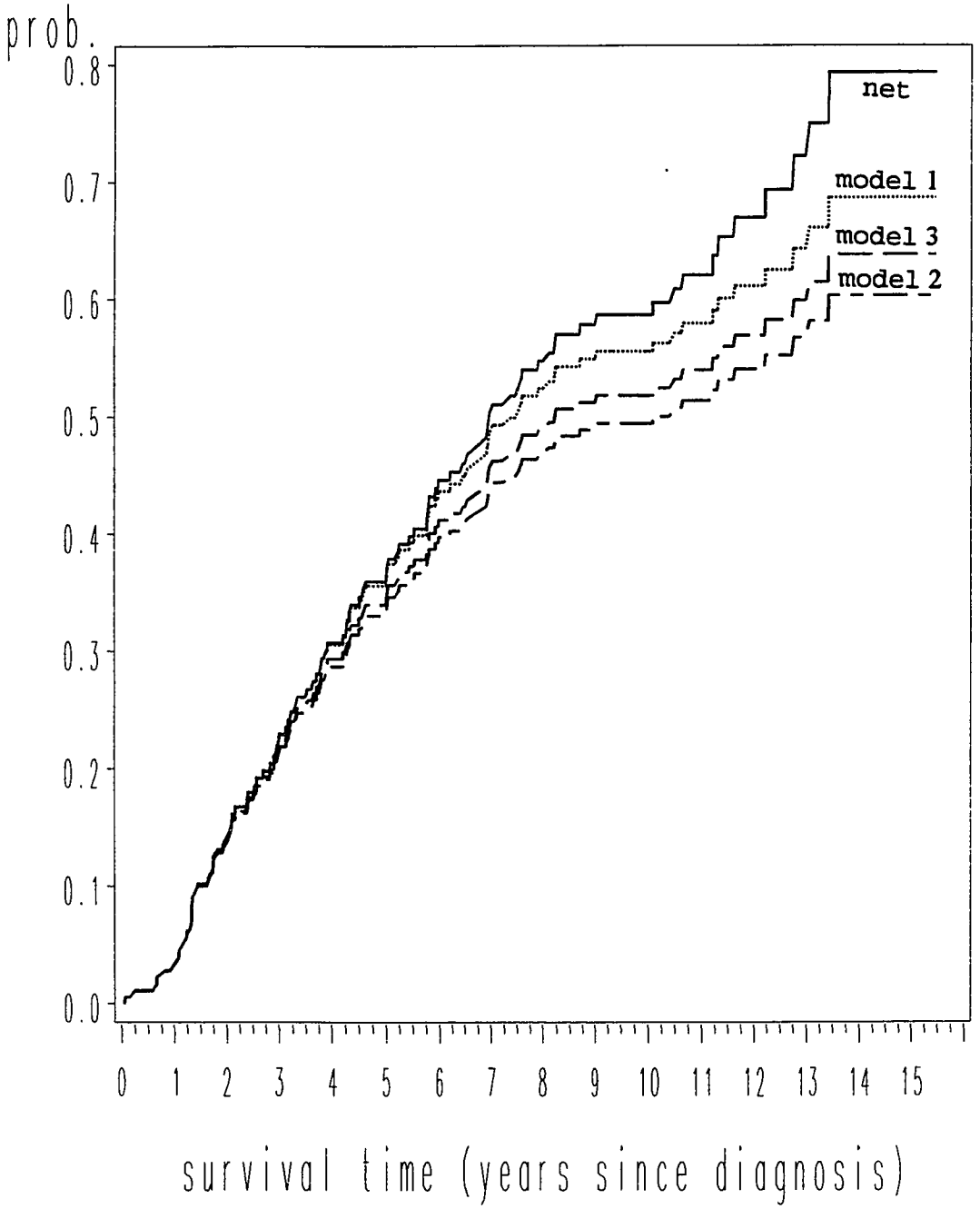


Figure 7. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;66, stage;distant.

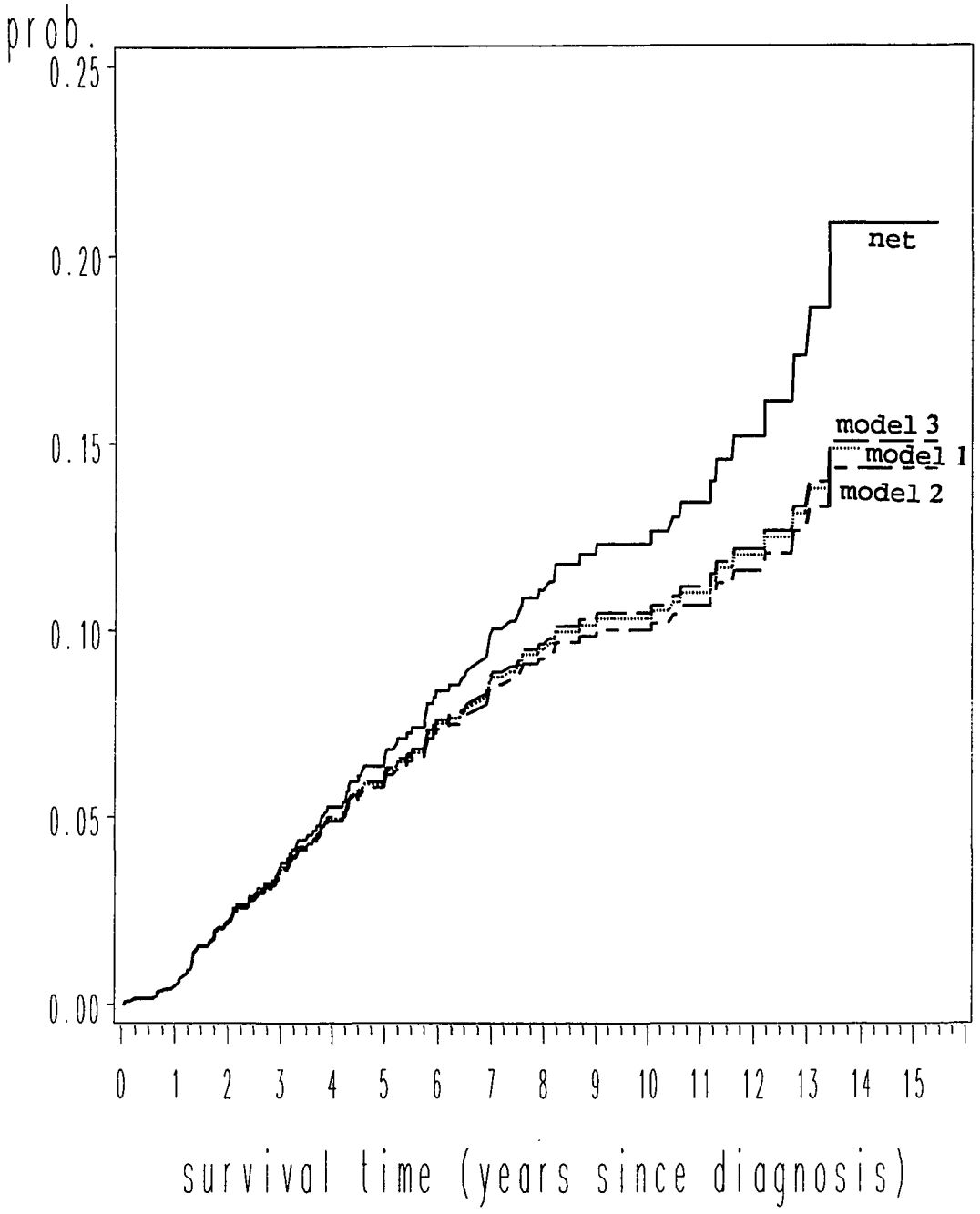


Figure 8. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;71, stage;localized.

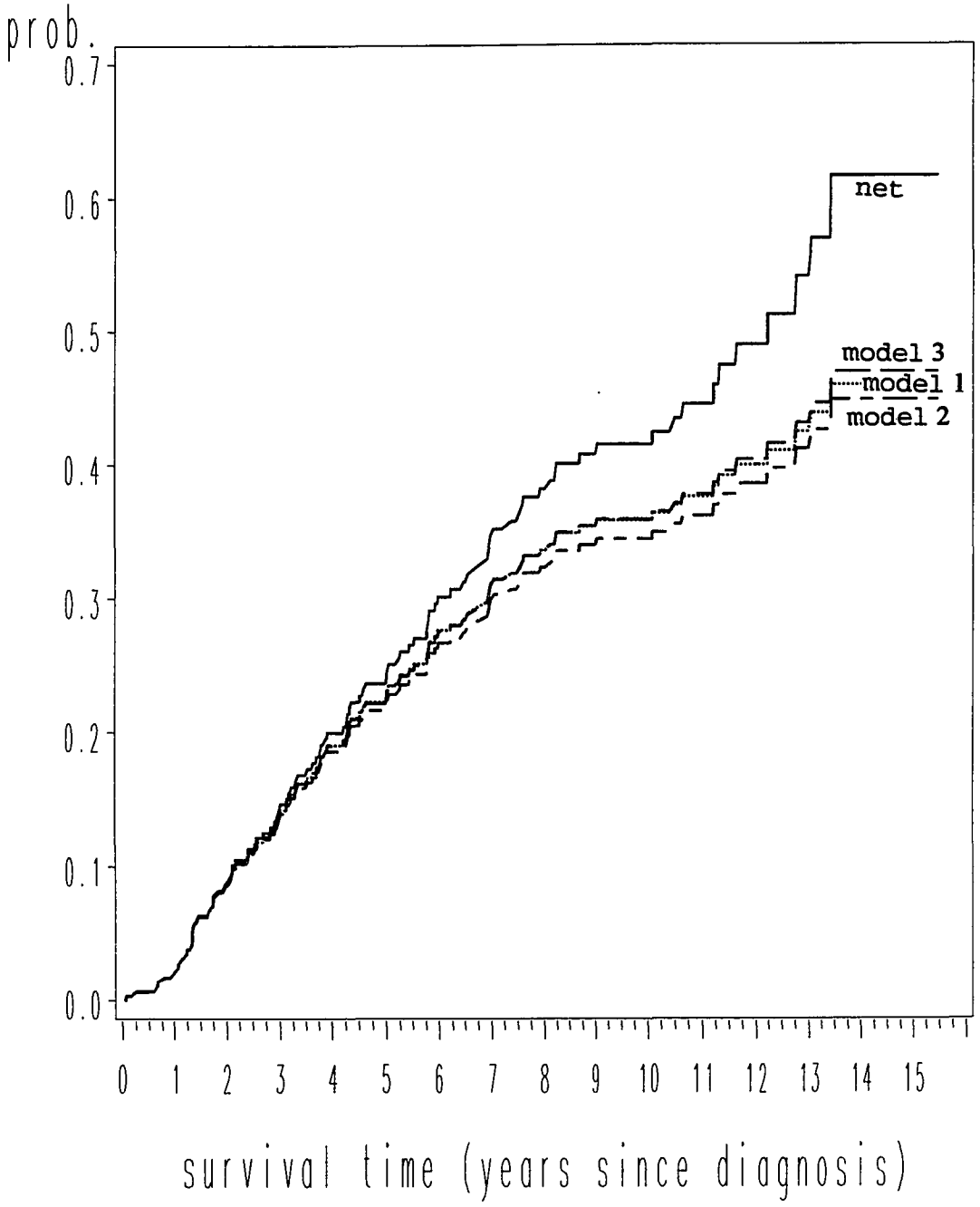


Figure 9. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;71, stage;regional.

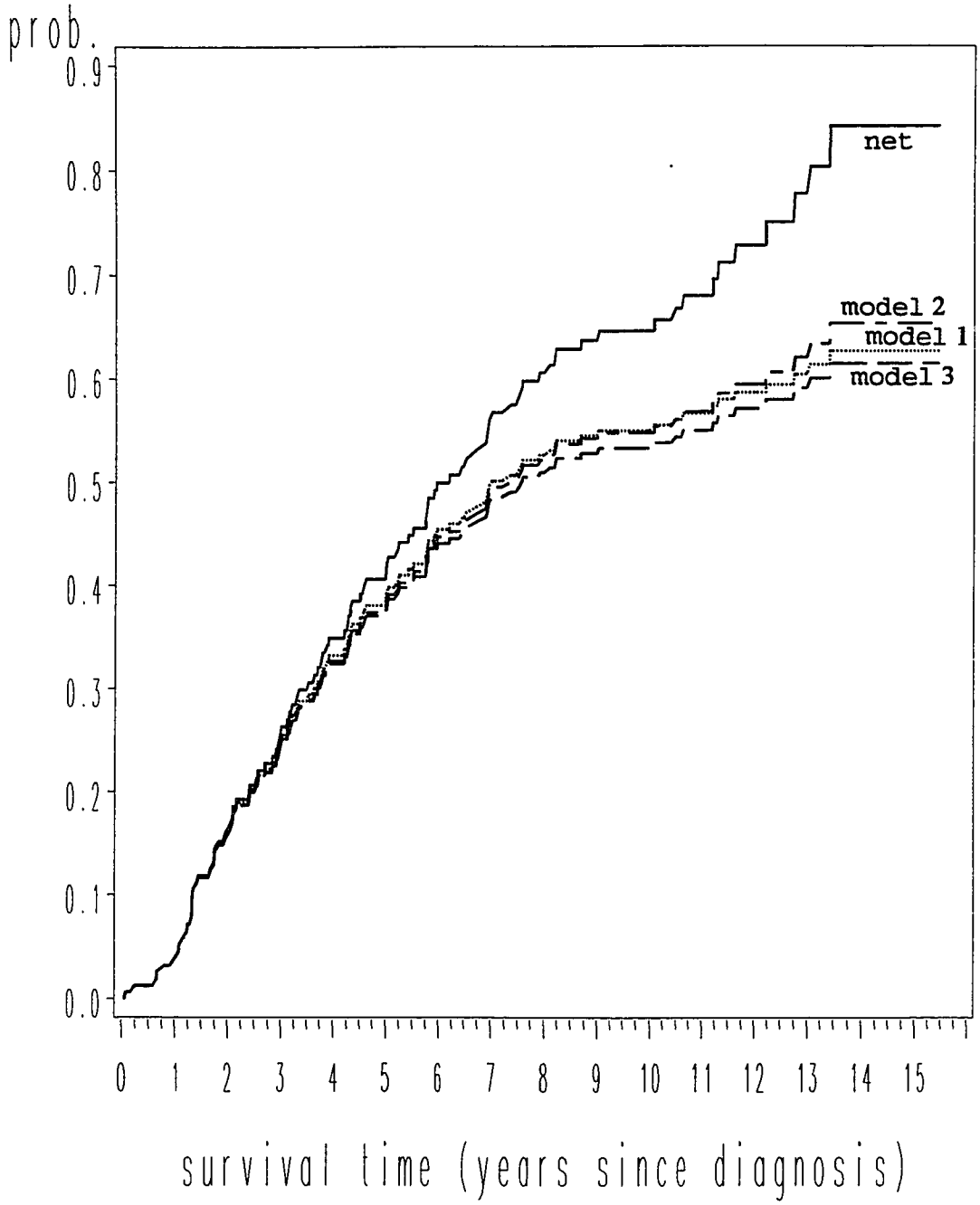


Figure 10. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;71, stage;distant.

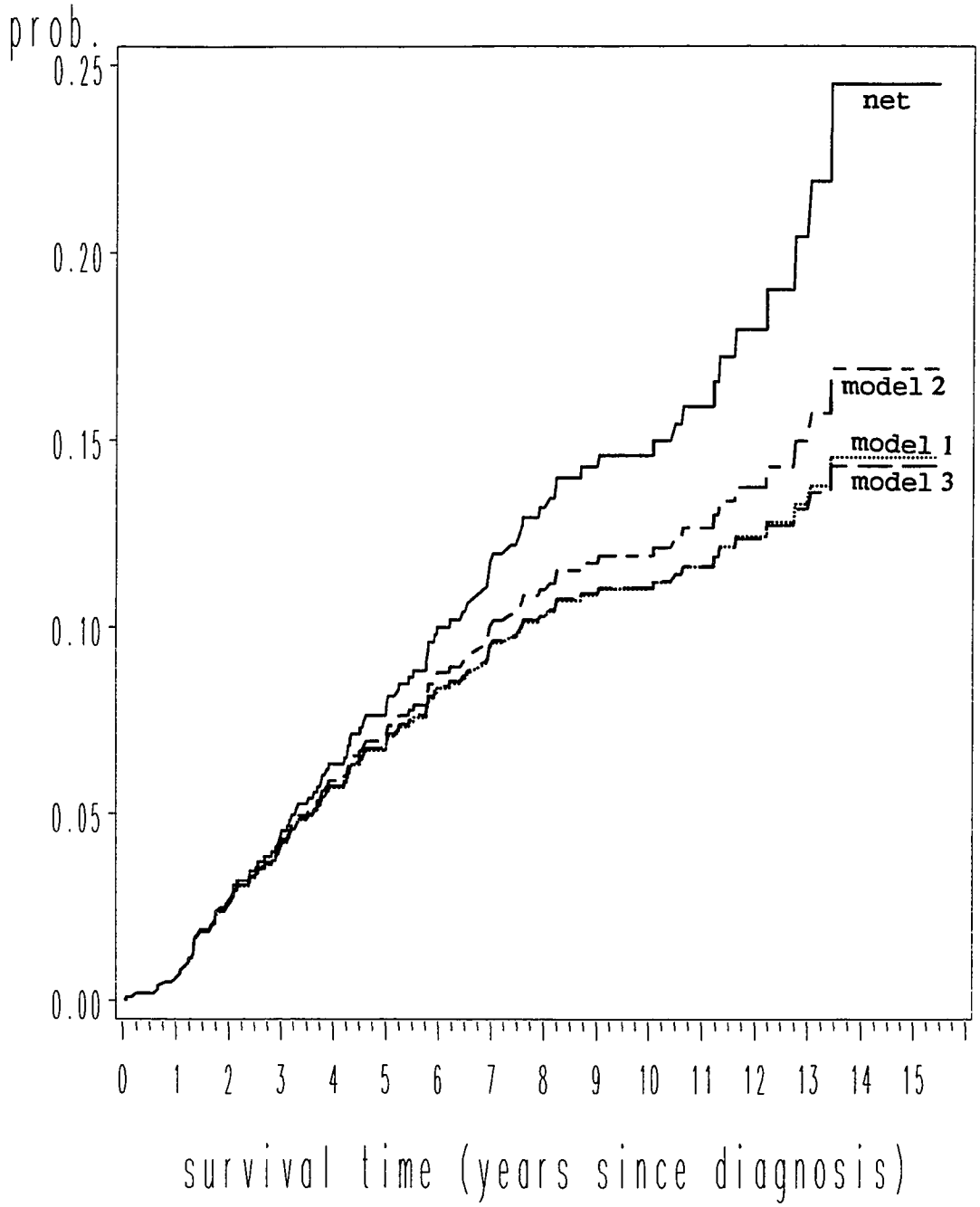


Figure 11. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;77, stage;localized.

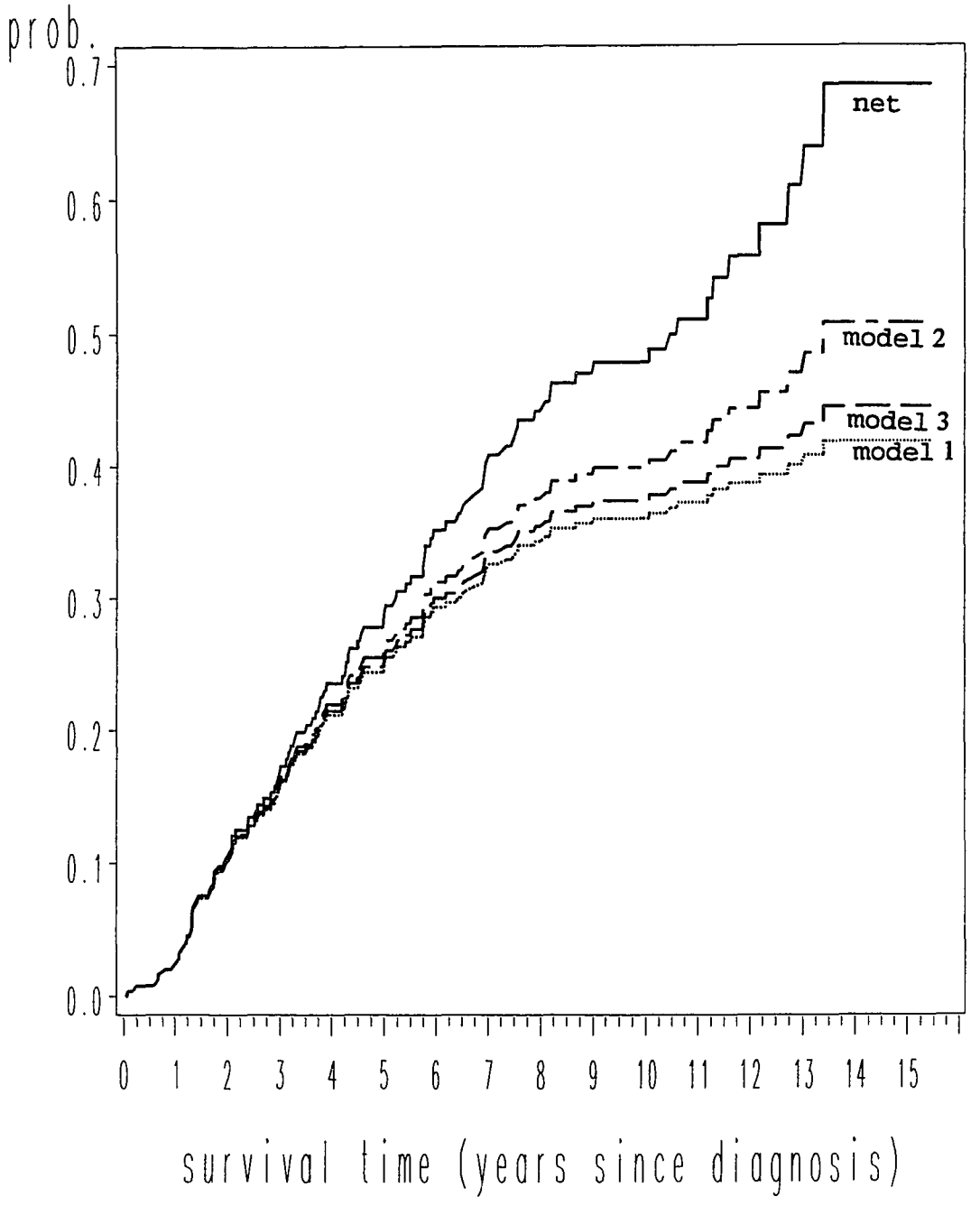


Figure 12. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;77, stage;regional.

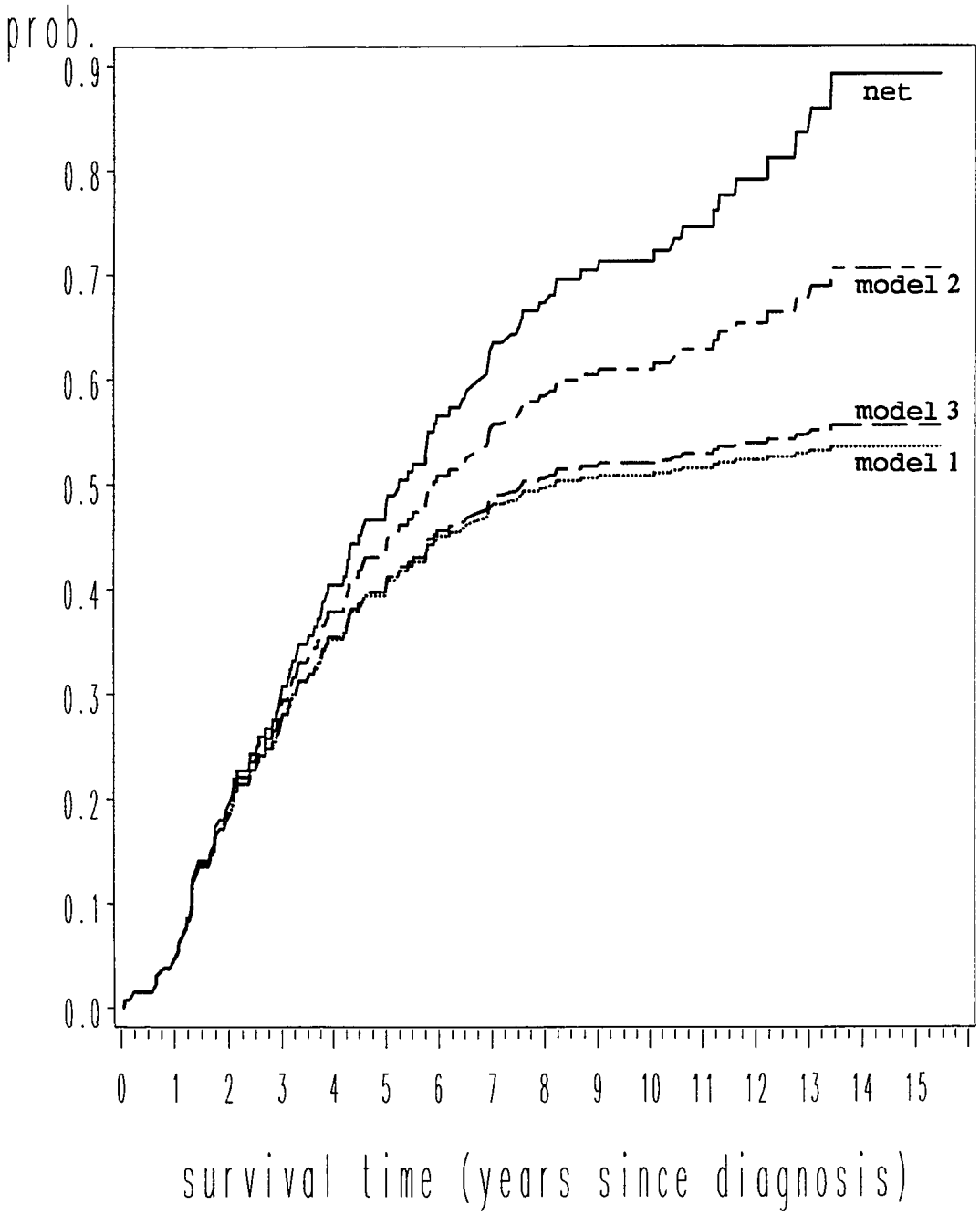


Figure 13. Comparison of estimates of net probability and models 1-3 for crude probability. Prostate cancer patients in Hawaii, age;77, stage;distant.

Table 11

The estimate ratio of net probability and crude probability of death from cause 1. Four time points, by age and stage, Prostate cancer patients.*

	Type of model in crude probability		
Time**	Model 1	Model 2	Model 3
		Age=66, Stage=localized	
3.71	1.037	1.071	1.031
7.66	1.103	1.196	1.089
11.11	1.139	1.260	1.122
15.45	1.258	1.459	1.232
		Age=71, Stage=localized	
3.71	1.058	1.070	1.050
7.66	1.162	1.195	1.147
11.11	1.218	1.259	1.201
15.45	1.404	1.455	1.386
		Age=77, Stage=localized	
3.71	1.098	1.070	1.091
7.66	1.275	1.194	1.269
11.11	1.369	1.257	1.369
15.45	1.684	1.449	1.714

$$* \ \frac{\textit{Estimate of Net probability}}{\textit{Estimate of Crude probability from Model } 1, 2, \textit{ or } 3}$$

** Quartiles of survival time.

Table 11 (Continued)

The estimate ratio of net probability and crude probability of death from cause 1. For four time points, by age and stage of prostate cancer patients.*

	Type of model in crude probability		
Time**	Model 1	Model 2	Model 3
		Age=66, Stage=regional	
3.71	1.014	1.067	1.028
7.66	1.058	1.181	1.082
11.11	1.086	1.236	1.110
15.45	1.174	1.388	1.192
		Age=71, Stage=regional	
3.71	1.042	1.066	1.047
7.66	1.132	1.178	1.133
11.11	1.183	1.231	1.178
15.45	1.339	1.373	1.312
		Age=77, Stage=regional	
3.71	1.098	1.065	1.085
7.66	1.279	1.173	1.240
11.11	1.371	1.223	1.317
15.45	1.642	1.353	1.545

$$* \ \frac{\textit{Estimate of Net probability}}{\textit{Estimate of Crude probability from Model } 1, 2, \textit{ or } 3}.$$

** Quartiles of survival time.

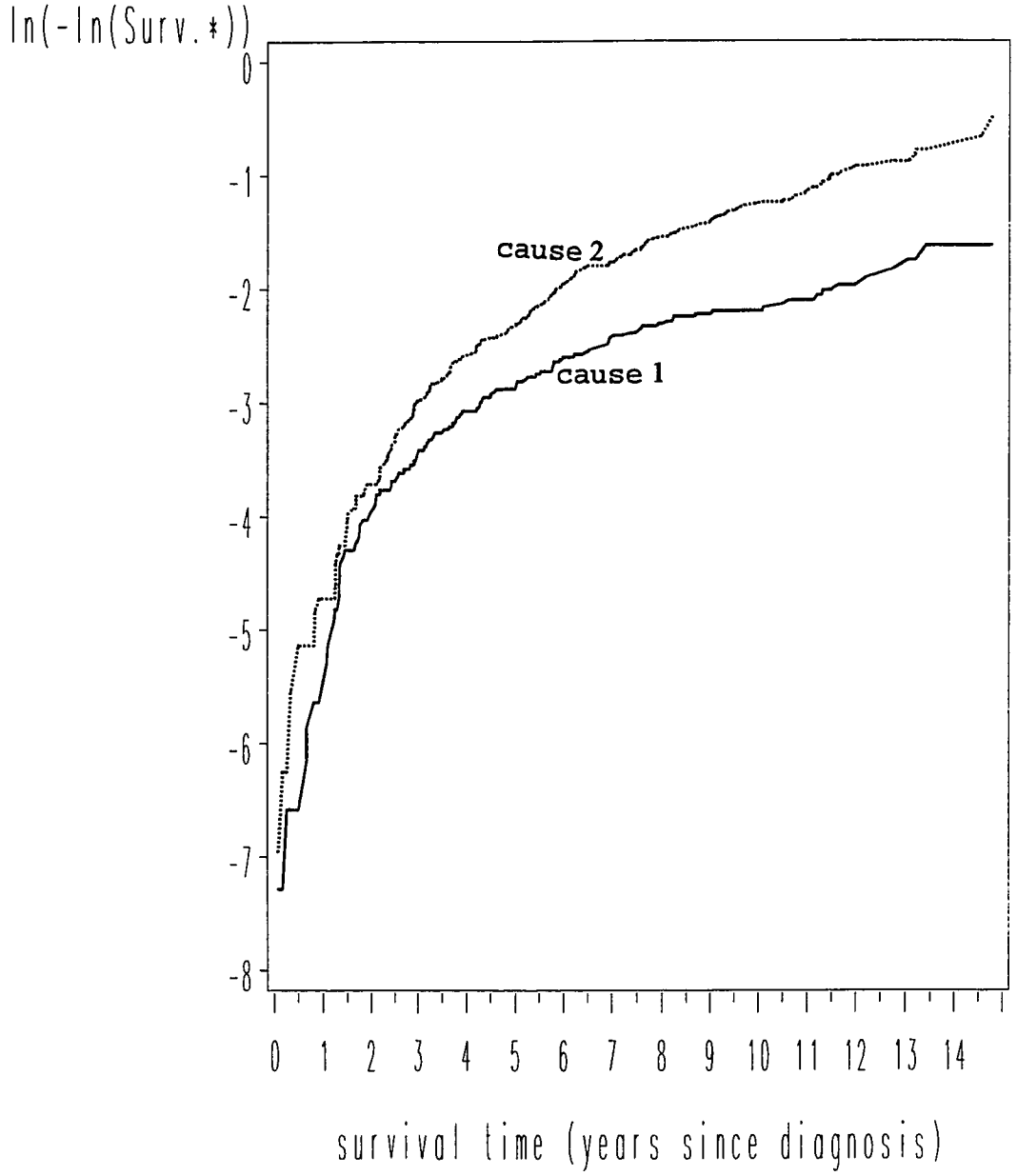
Table 11 (Continued)

The estimate ratio of net probability and crude probability of death from cause 1. For four time points, by age and stage of prostate cancer patients.*

	Type of model in crude probability		
Time**	Model 1	Model 2	Model 3
		Age=66, Stage=distant	
3.71	1.002	1.063	1.041
7.66	1.042	1.164	1.114
11.11	1.071	1.208	1.150
15.45	1.154	1.314	1.242
		Age=71, Stage=distant	
3.71	1.044	1.062	1.068
7.66	1.147	1.158	1.184
11.11	1.201	1.199	1.238
15.45	1.346	1.292	1.374
		Age=77, Stage=distant	
3.71	1.128	1.060	1.125
7.66	1.349	1.150	1.324
11.11	1.445	1.186	1.408
15.45	1.663	1.263	1.602

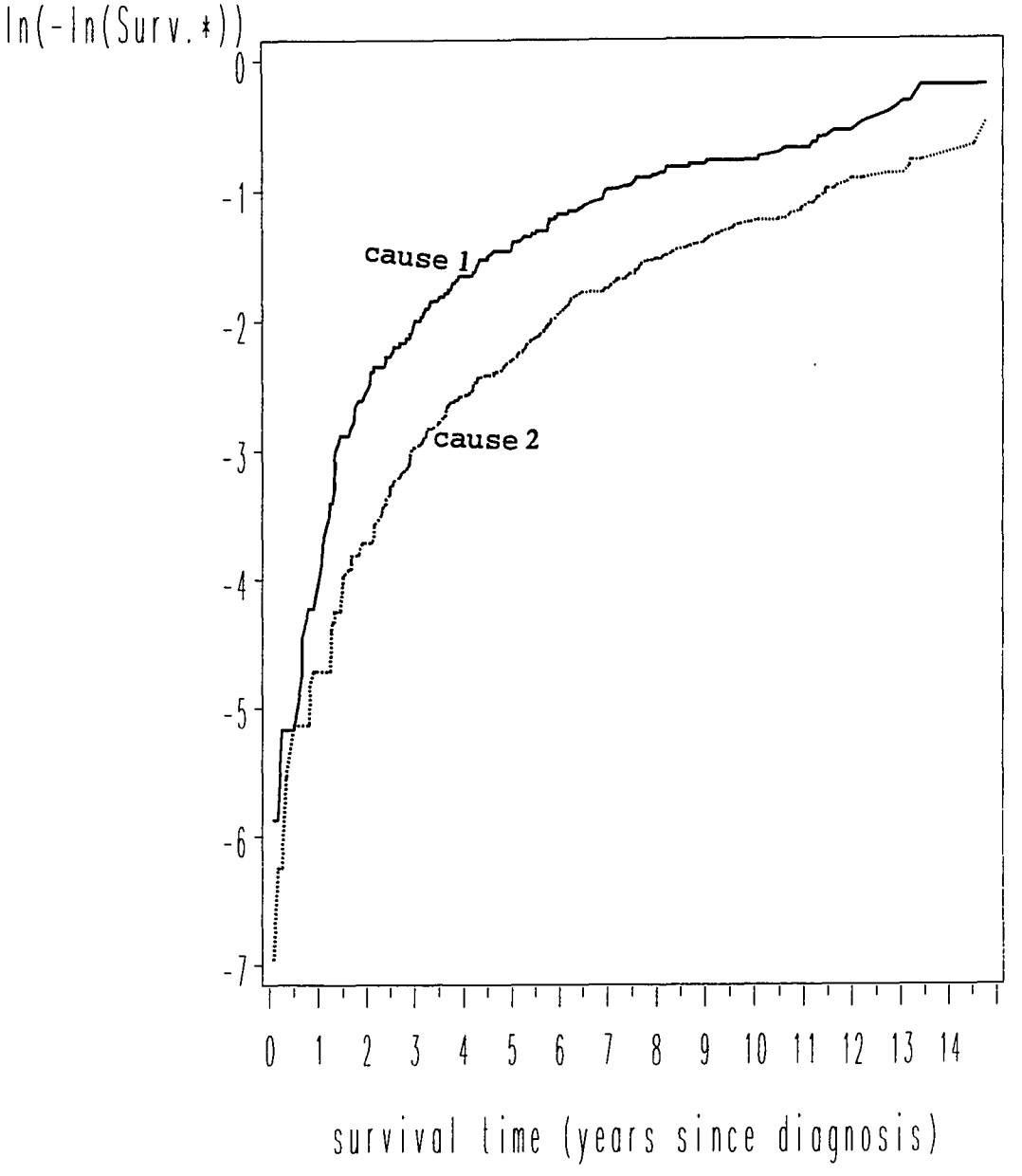
* $\frac{\textit{Estimate of Net probability}}{\textit{Estimate of Crude probability from Model 1,2 or 3}}$.

** Quartiles of survival time.



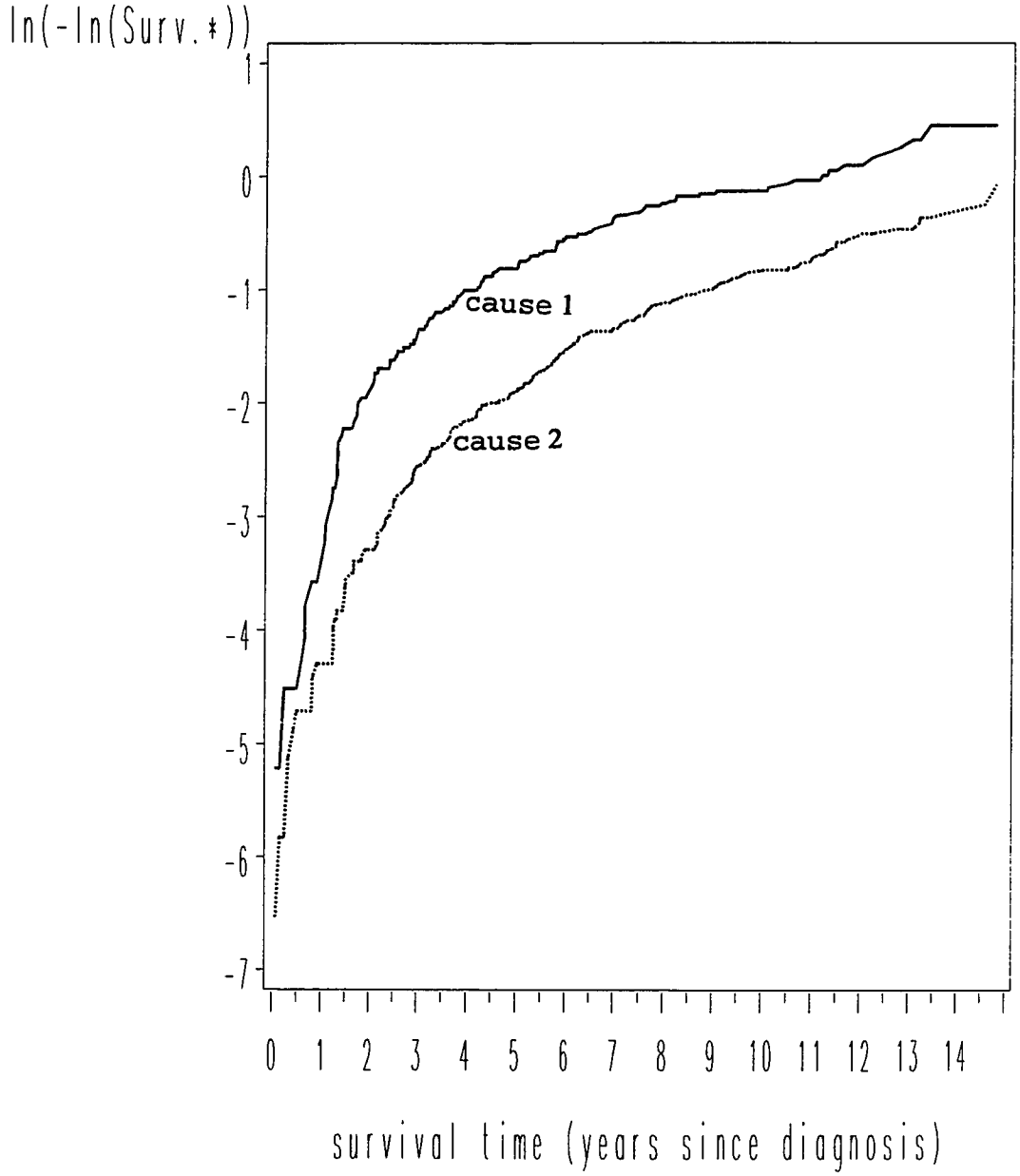
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 14. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 66, stage localized



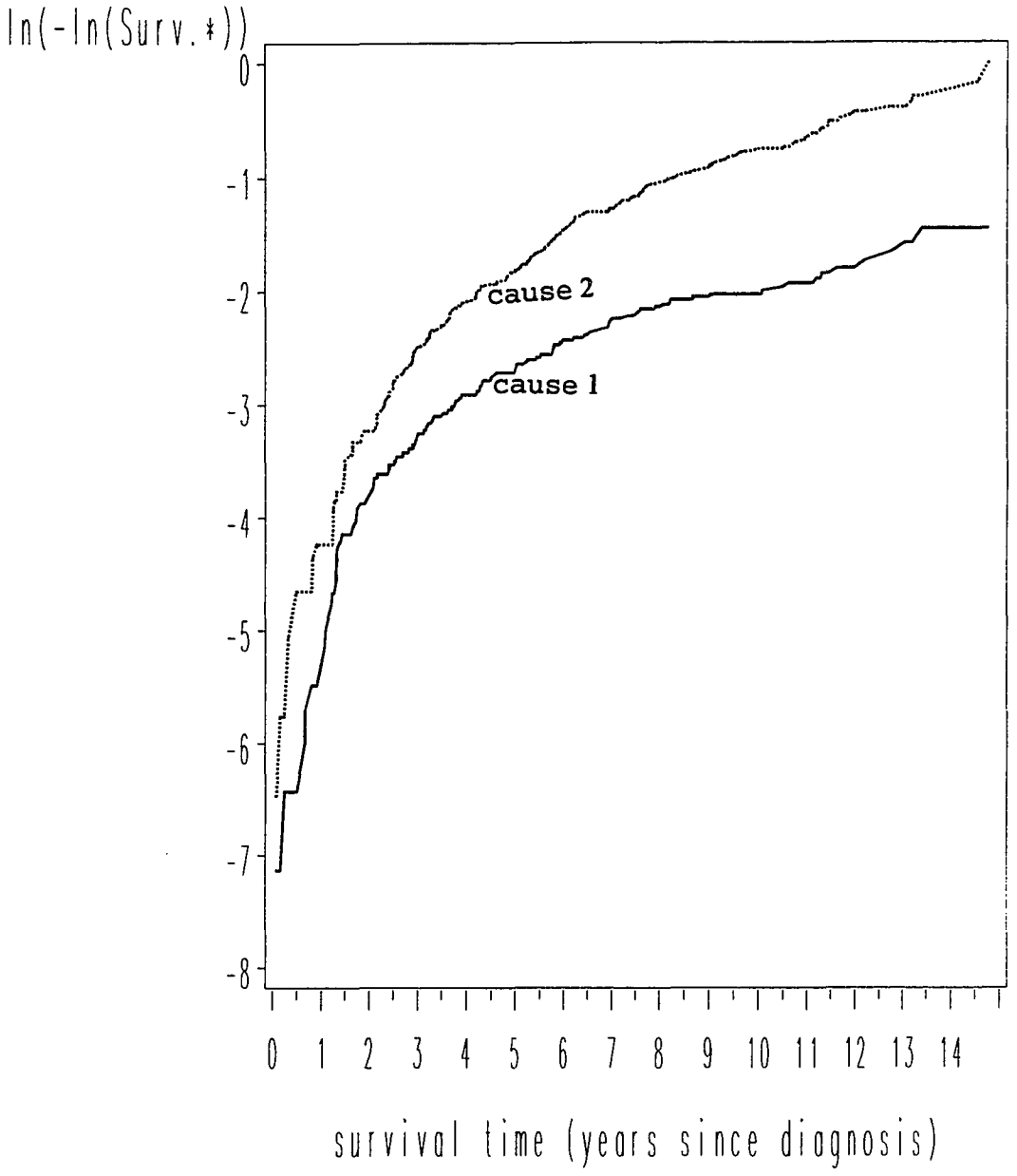
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 15. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 66, stage regional



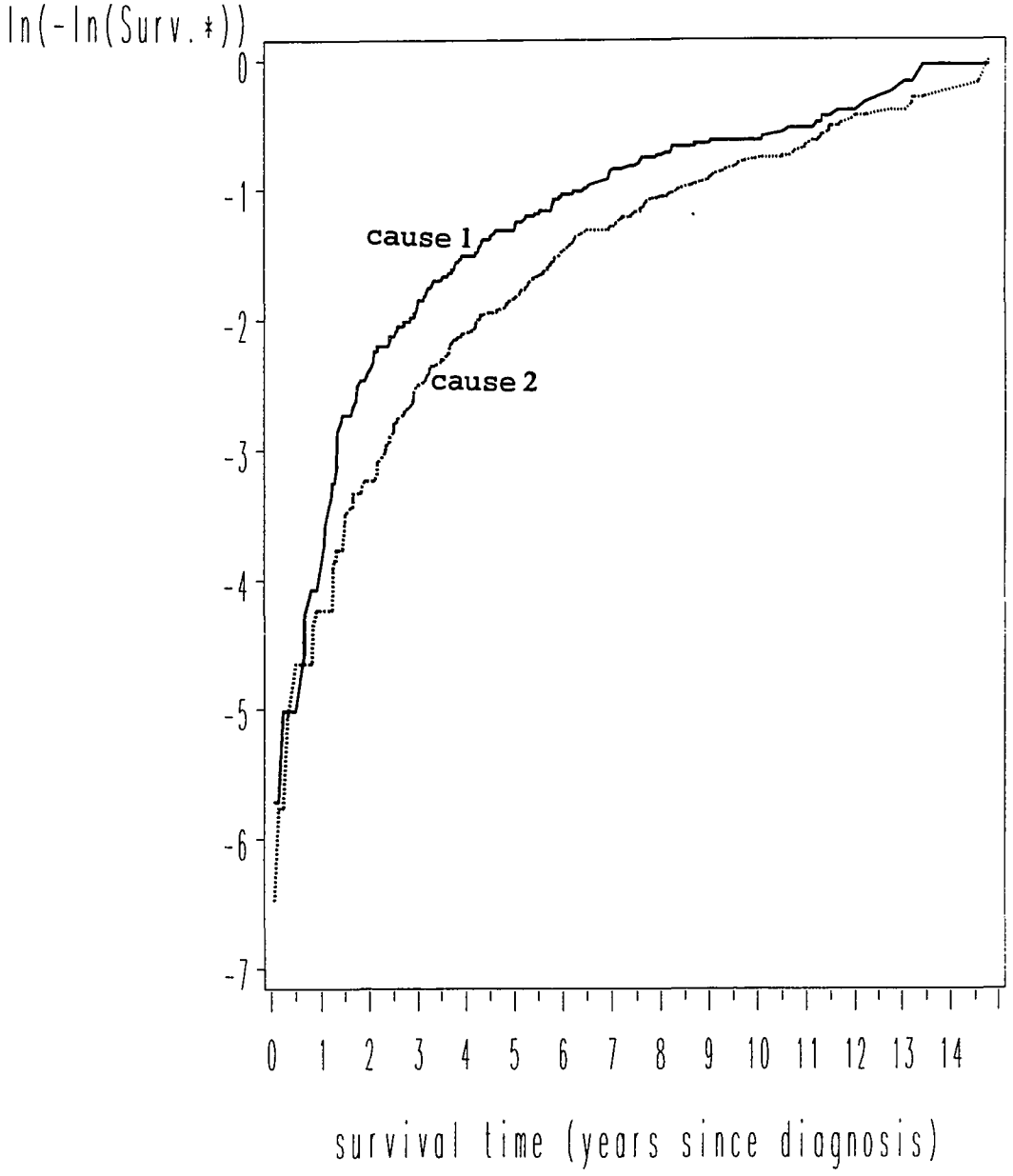
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 16. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 66, stage distant



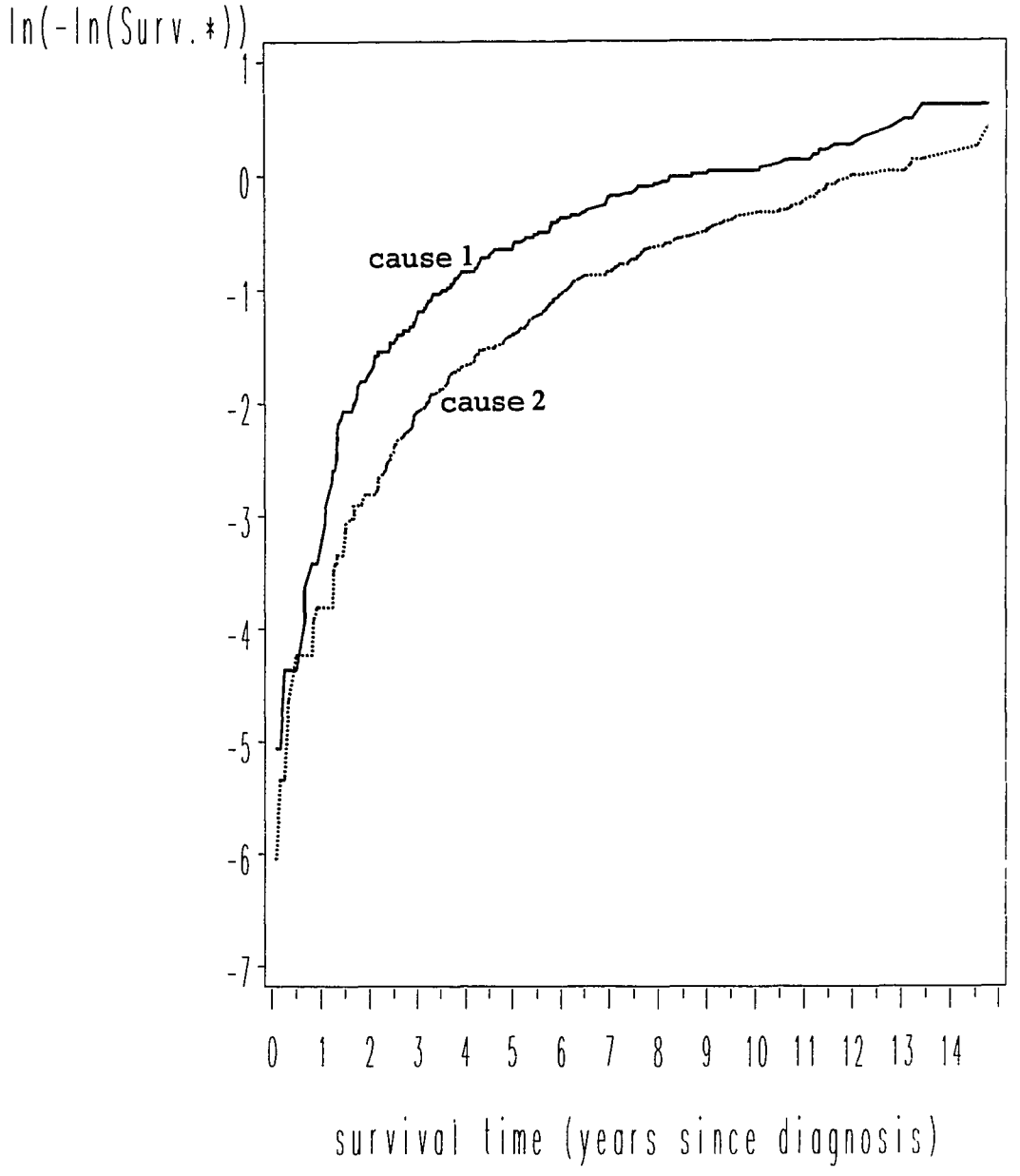
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 17. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 71, stage localized



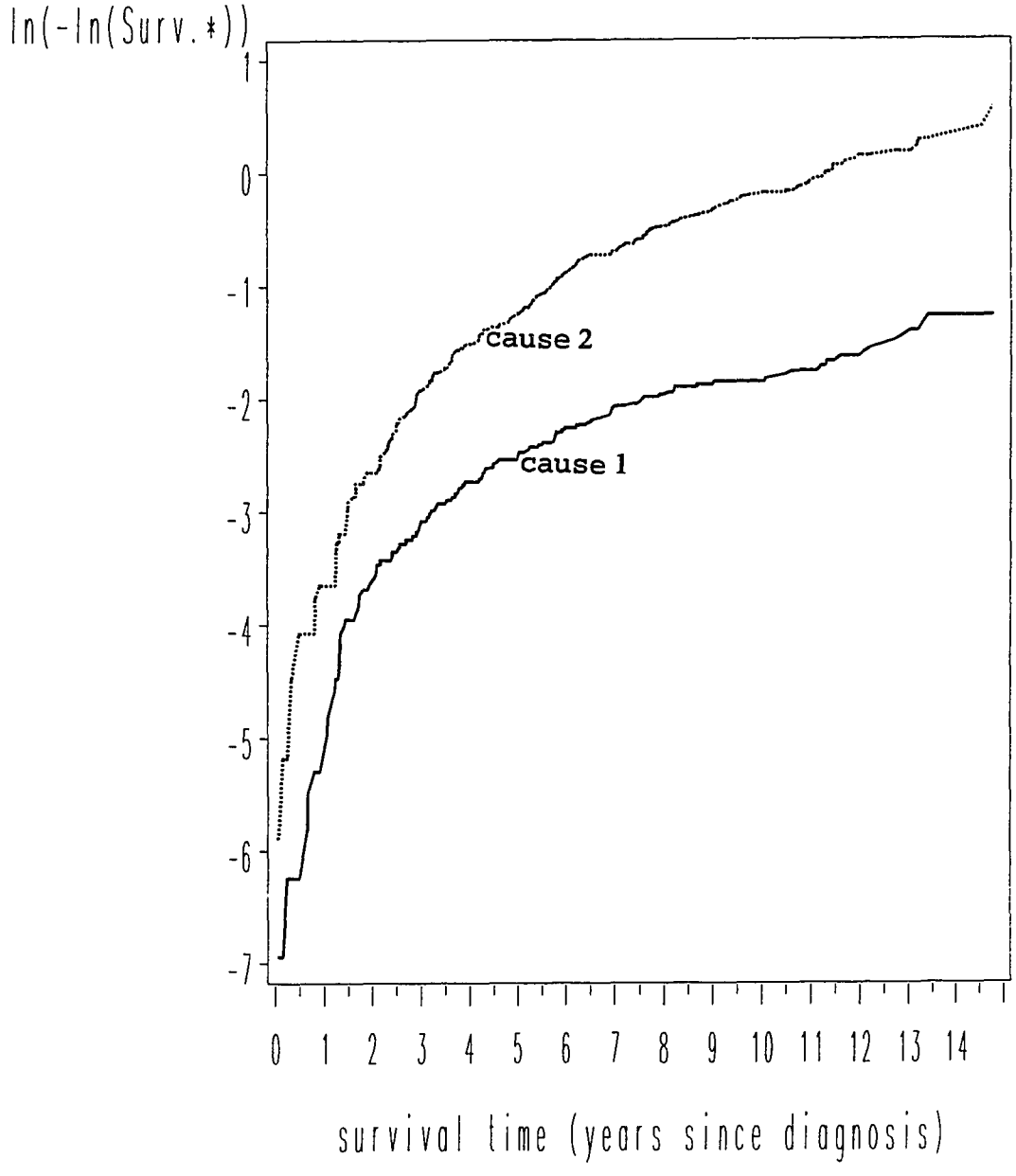
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 18. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 71, stage regional



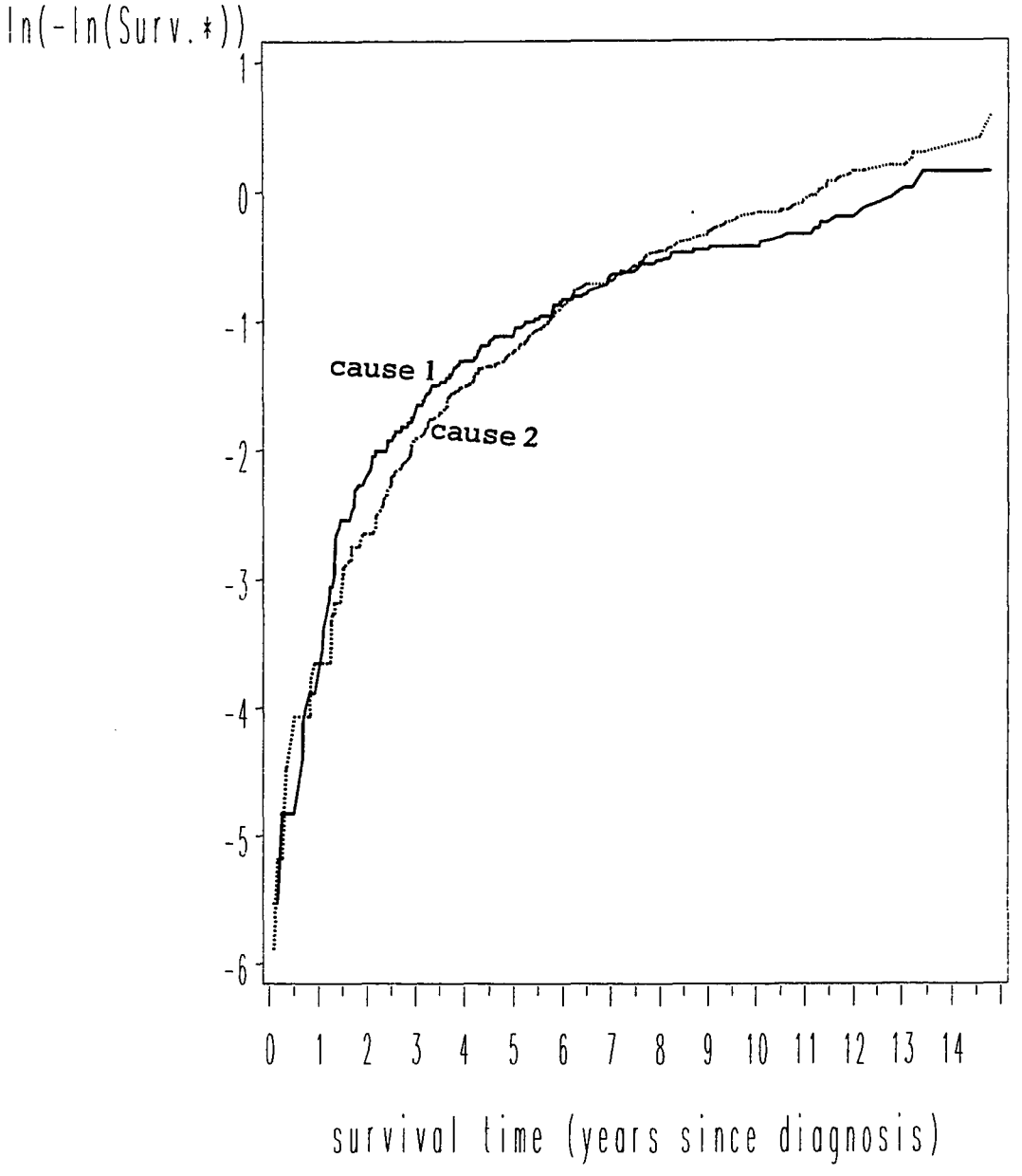
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 19. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 71, stage distant



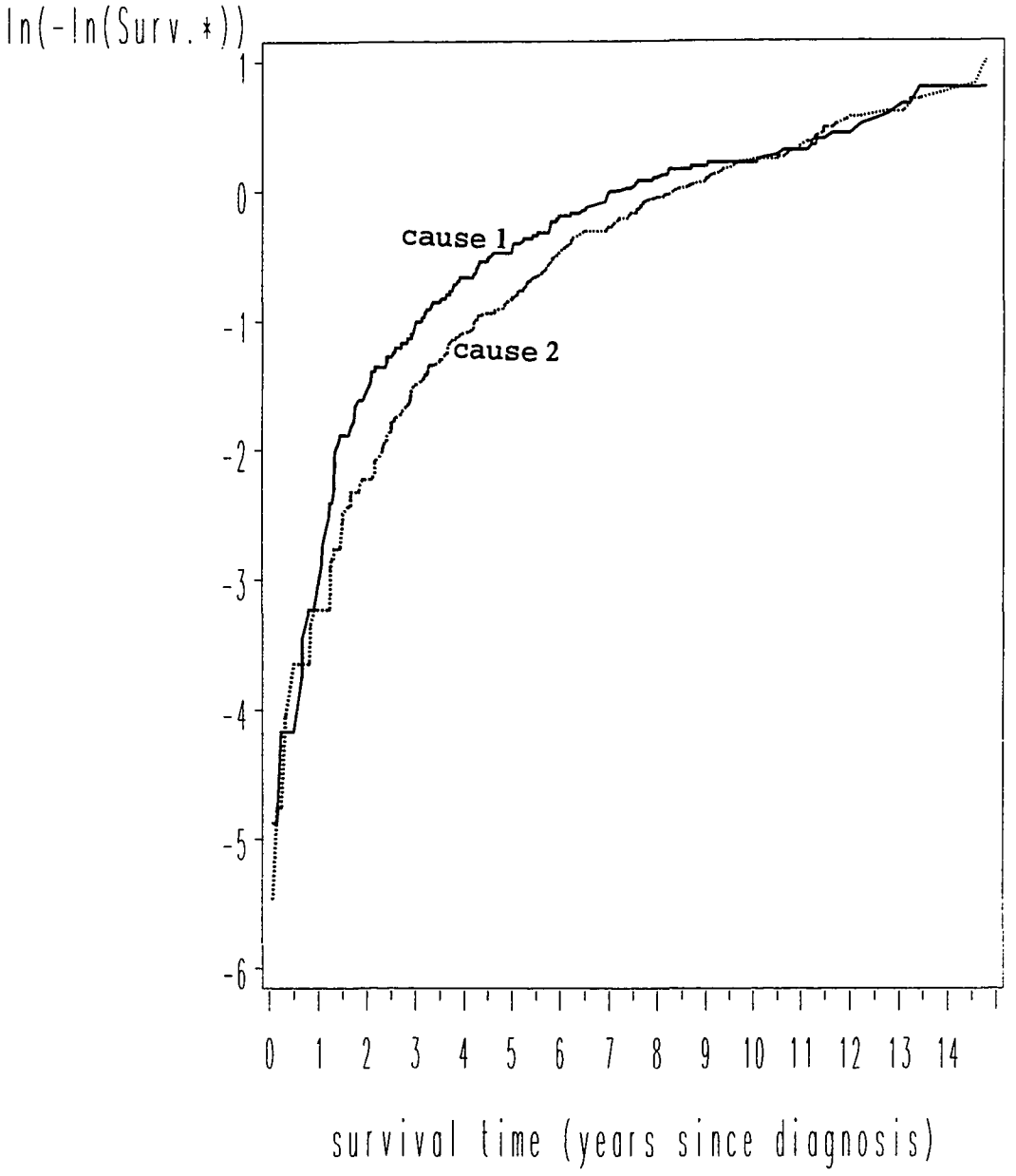
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 20. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 77, stage localized



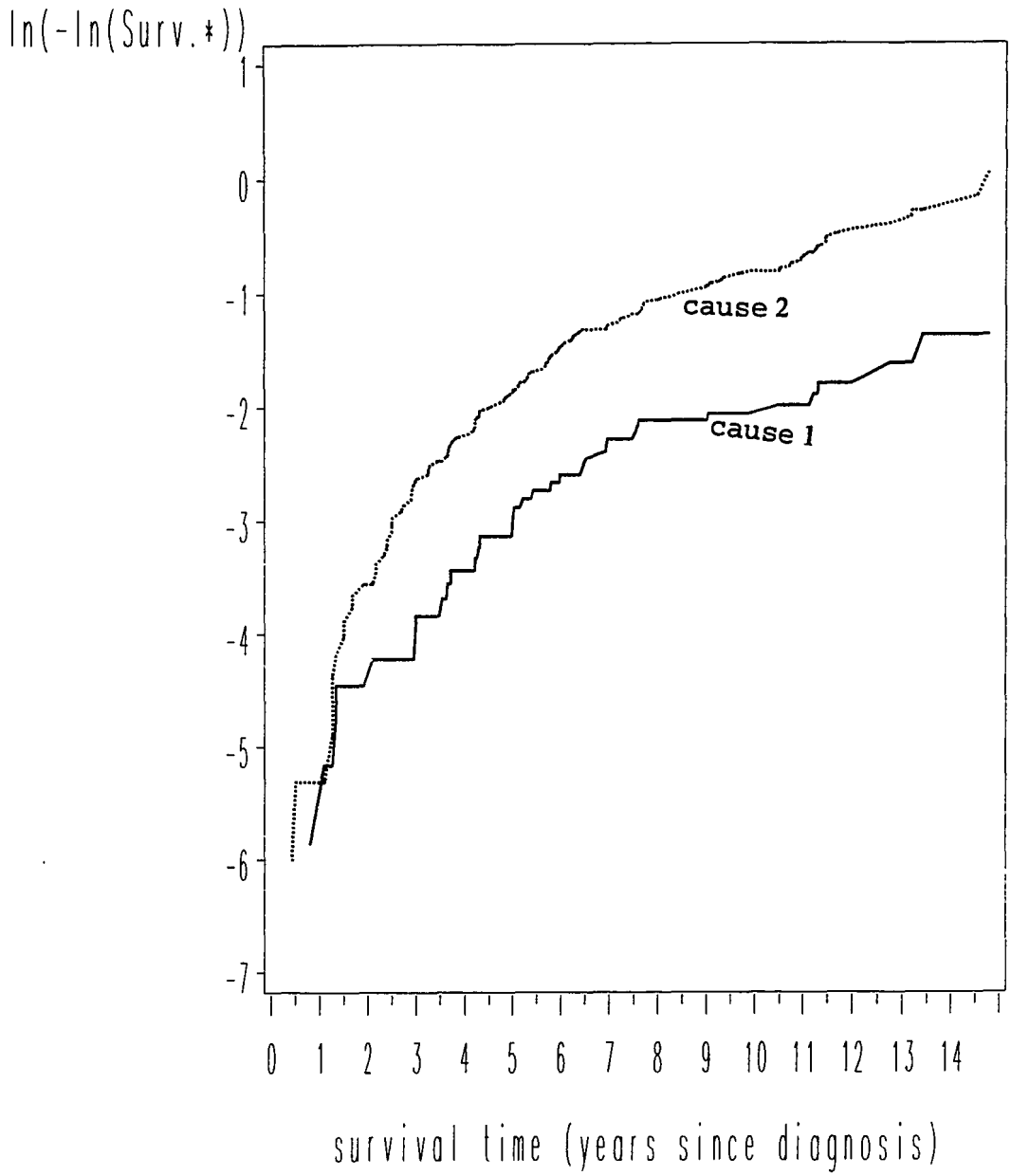
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 21. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 77, stage regional



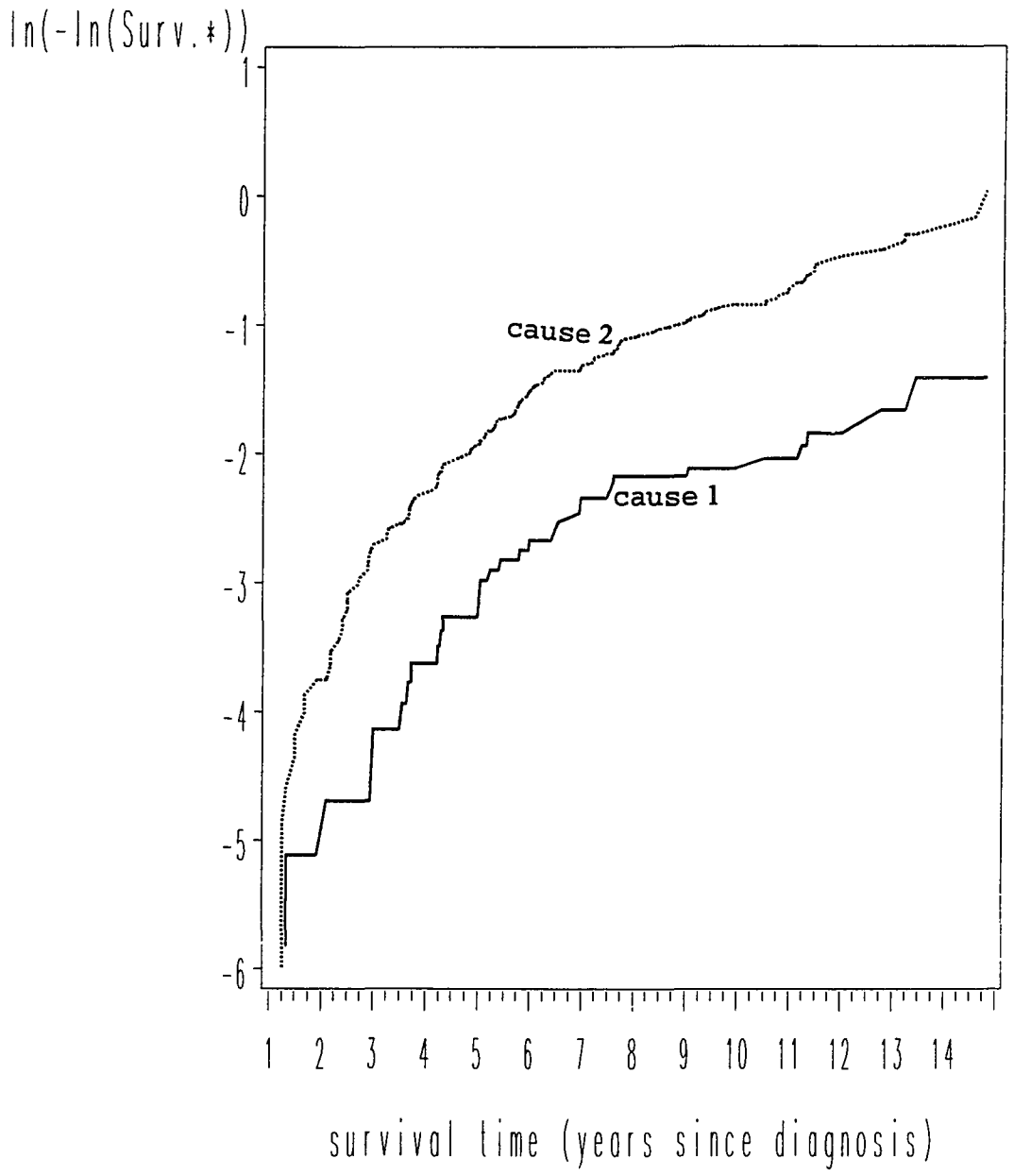
* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given values of covariate

Figure 22. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2), age 77, stage distant



* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given value of covariate

Figure 23. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2) in stratum stage localized, age 71



* Kaplan-Meier estimates of cause of interest by treating extraneous causes as censored, for given value of covariate

igure 24. Graphical examination of proportionality assumption for prostate cancer (cause 1) and other causes (cause 2) in stratum stage localized where the survival time exceeds 1.2 yrs

Table 12

Estimates of coeffici s (β,β_j) of age under proportional hazards models for overall (prostate cancer and other causes) and cause-specific hazards by stage and appropriate methods to estimate crude probability of death from prostate cancer. Standard errors are given in parentheses.

		Cause-specific		Appropriate
Stage	Overall	Prostate cancer	Other causes	Methods
Localized	0.100(0.012)*	0.081(0.025)*	0.105(0.014)*	**
				1. ***Non-parametric for survival time < 1.2 years (Time zone one)
				2. Model 1 & Model 3 for survival time ≥ 1.2 years (Time zone two)
Regional	0.040(0.023)	-0.003(0.032)	0.085(0.033)*	****Modification of Model 2
Distant	0.034(0.016)*	0.013(0.020)	0.069(0.028)*	****Modification of Model 2

* $p < 0.05$

** Two time zones were made because of a crossing around survival time 1.2 years in checking for the proportionality assumption for prostate cancer and other causes in stage 'Localized'.

*** $\hat{F}_j(t) = \sum_{\{i:t(i)\le t\}} \frac{d_{ji}}{n_i} \hat{S}(t_{(i-1)})$

**** $\hat{F}_1(t,z) = \sum_{\{i:t(i)\le t\}} \hat{h}_1(t)\hat{G}_1(t)\hat{G}_{02}(t)^{\exp(z\hat{\beta}_2)}$.

Table 13

Estimates of crude probability of death from prostate cancer by alternative method. For four time points, by age and stage, prostate cancer patients.

Time*	'Localized' model (1)**	'Localized' model (3)**	'Regional'	'Distant'
		Age=66		
3.71	.0243	.0243	.1419	.3561
7.66	.0743	.0743	.3131	.5043
11.11	.0824	.0824	.3831	.5672
15.45	.1293	.1293	.5001	.6265
		Age=71		
3.71	.0319	.0319	.1381	.3469
7.66	.0978	.0979	.2970	.4828
11.11	.1075	.1076	.3575	.5346
15.45	.1576	.1578	.4503	.5803
		Age=77		
3.71	.0447	.0447	.1311	.3315
7.66	.1283	.1285	.2691	.4482
11.11	.1383	.1385	.3149	.4853
15.45	.1792	.1793	.3744	.5143

* Quartiles of survival time.

** Denotes method of estimation of crude probability for time zone two in 'localized'. The estimates of crude probability for time zone one are added to those for time zone two.

4 strata

'Localized' survival time < 1.2 years (Time zone one)
survival time ≥ 1.2 years (Time zone two)

'Regional'

'Distant'

Table 14

Estimates of crude probability of death from other causes by alternative method. For four time points, by age and stage, prostate cancer patients.

Time*	'Localized' model (1)**	'Localized' model (3)**	'Regional'	'Distant'
		Age=66		
3.71	.0588	.0589	.0981	.1118
7.66	.1703	.1703	.1721	.1914
11.11	.2494	.2492	.2172	.2584
15.45	.4288	.4288	.2760	.
		Age=71		
3.71	.0934	.0935	.1464	.1537
7.66	.2647	.2649	.2505	.2558
11.11	.3743	.3745	.3103	.3340
15.45	.5877	.5887	.3857	.
		Age=77		
3.71	.1636	.1636	.2328	.2222
7.66	.4266	.4272	.3802	.3530
11.11	.5616	.5626	.4560	.4380
15.45	.7549	.7569	.5455	.

* Quartiles of survival time.

** Denotes method of estimation of crude probability for time zone two in 'localized'. The estimates of crude probability for time zone one are added to those for time zone two.

4 strata

'Localized' survival time < 1.2 years (Time zone one)
survival time ≥ 1.2 years (Time zone two)

'Regional'

'Distant'

Table 15

The estimate ratio of net probability and model 3 for crude probability for cause 1 and cause 2. Four time points, by age** where time zone two of stage localized, prostate cancer patients.*

Cause time***	Cause 1*	Cause 2**
	age=66	
3.71	1.031	1.006
7.66	1.107	1.031
11.11	1.125	1.048
15.45	1.250	1.078
	age=71	
3.71	1.053	1.009
7.66	1.187	1.044
11.11	1.217	1.068
15.45	1.439	1.107
	age=77	
3.71	1.103	1.013
7.66	1.369	1.065
11.11	1.429	1.099
15.45	1.873	1.144

* $\dfrac{\textit{Estimate of net probability}}{\textit{Estimate of Model 3 for crude probability}}$.

** $\hat{\beta}_1 = 0.0776, \quad \hat{\beta}_2 = 0.1063$.

*** Quartiles of survival time.

cause 1* cause of interest; prostate cancer.

cause 2** cause of interest; other causes.

CHAPTER 5

SUMMARY AND DISCUSSION

When infectious diseases were the main killers, elimination of their effects on mortality rates was possible. However, in modern society in which chronic disease are the major causes of death, elimination of the disease effect is no longer relevant in estimating mortality rates (4,8). Therefore, studies on cause-specific failure probabilities are appropriate under the practical situation of other causes acting simultaneously without imposing the independent competing risks assumption rather than under the impractical assumption of elimination of other causes.

The crude probability is smaller than the net probability when competing risks exist because the probability of survival from other causes should be accounted for to obtain the crude probability of death from cause of interest. From the analysis of simulated data, it was found that the difference between the estimate of net probability and that of crude probability becomes larger as the proportion censored (due to loss-to-follow-up and

termination of study) gets smaller. If the proportion censored is held constant, the difference becomes greater as the proportion of deaths from the cause of interest (compared to other causes) gets smaller. These results are consistent with those of the study by Gaynor et al. (12) in which they used clinical oncology data.

Quantities from the proportional hazards model are used to estimate the CSFP when covariate effects are considered. Fitting the model is performed simply by treating the patients who fail due to other causes as censored observations (1,2). Overall, disease stage at diagnosis affects risk of death by prostate cancer more than age does, while age affects the hazard of other causes more than stage does. The estimate of net probability is greater than that of crude probability when the effects of covariates are considered. When the coefficient of a covariate is positively related to the hazards of other causes, the estimate of net probability is increased over that of crude probability according to the increasing value of the covariate. It is also noticed that the more the association

of a covariate is related to other causes, the larger the difference is.

If the proportionality assumption of Cox's model is violated for an explanatory variable, stratification for the variable is desirable and CSFP is estimated for each stratum.

If the proportional hazards model fits selected causes, strata, or time zones, a combination of non-parametric and semiparametric hazard and survival functions can be used to estimate the CSFP.

Model 1 (The crude probability defined as the integral of the product of the overall survival function times the cause-specific hazard at time t with covariate effects over $(0,t]$) is appropriate for cause-specific failure probability with covariate effects. It does not require the restriction of no tied failure times involving different causes of death. However, if the covariates do not explain or partly explain one of the causes of death in the proportional hazard model, either model 2 (The crude probability under the assumption that the hazard of cause 2 does not depend on covariates. It is defined the same as

the model 1, but overall survival, $S(t;z)$, in the model is composed of multiplication of survival of cause 1, $G_1(t;z)$, and survival of cause 2, $G_2(t)$.) or model 3 (The crude probability under the situation that the hazard of cause 2 depends on covariates as the hazard of cause 1 does. It is defined the same as model 1, but overall survival, $S(t;z)$, in the model is composed of multiplication of survival of cause 1, $G_1(t,z)$, and survival of cause 2, $G_2(t;z)$.) can be used as alternative methods. In using these two models, tied failure times involving different causes of death are not allowed.

Assessing the importance of competing risks over a period of time has long been a concern of demographers, vital statisticians, and actuaries, the last preferring the term multiple decrements to competing risks (5).

Competing risk survival analysis is the generalization of ordinary survival analysis in which each item under testing may fail for just one of a number of different causes with or without censoring (44). To choose a method of analysis for cause-specific failure probability depends

on the situation. If physically distinct causes of failure are available, like in industrial reliability testing, the methods that assume a single cause is appropriate to obtain the cause-specific failure probability just by treating extraneous causes as censored observations. In this situation, the survival function of the cause of interest is obtained under the condition that failures due to other causes have been eliminated and the cause-specific hazard rate of interest, $h_1(t)$, remains unchanged (7,12).

While some researchers tend to use the net probability as the cause-specific failure probability by treating extraneous causes as censored observations, some researchers hesitate to use the assumption in a proportional hazards model on data which have more than one cause. As mentioned in the Introduction and Method sections, the estimation of a particular cause-specific hazard and tests of the covariate effects on that hazard, e.g., fitting of proportional hazards model, are performed simply by treating the patients who fail of extraneous causes as censored observations (1,2).

The concept of treating extraneous causes as censored observations cannot be extended to cause-specific failure probability studies on human mortality without unrealistic assumptions. Therefore, $\hat{F}_j(t)$ and $\hat{F}_j(t,z)$ ($\hat{F}_1(t,z)$ if $j = 1$, in models 2 and 3) are used for the analysis of cause-specific failure probability for homogeneous subjects and heterogeneous subjects with covariates, respectively. $\hat{F}_j(t)$ and $\hat{F}_j(t;z)$ have valid interpretations as probabilities even when competing risks are not independent (15).

As Prentice et al.(1) point out, the interpretation of cause-specific regression coefficients on the hazard function requires no assumptions concerning the interrelations among the causes of failure. Thus inference on the effects of treatment or exposure variables incorporated in z on specific types of failure can be made without introducing strong modeling assumptions. The interpretation of such effects is, however, restricted to the actual study conditions and there is no implication that the same regression estimates would prevail under a new set

of conditions where, for example, certain causes of failure have been eliminated (1).

As Benichou and Gail described in application of their absolute cause-specific failure risk (15), the cause-specific failure probability, $F_j(t)$, has a limitation for testing treatment effects. For example, if a cancer treatment increases the hazard of extraneous causes, h_2, but leaves the hazard of the cause of interest, h_1, unaffected, $F_1(t)$ will diminish in the treated group, yet overall survival is reduced and $G_1(t)$ is unchanged. Therefore, one should compare overall survival and estimates of the $G_1(t)$ and $G_2(t)$ in the treated and untreated groups, as is common practice. Here $G_2(t)=\exp\left\{-\int_0^t h_2(u)du\right\}$. If h_2 is not affected by treatment, the change in $F_1(t)$ is a more realistic gauge of the treatment benefit than a comparison of $G_1(t)$ curves. If both h_1 and h_2 are affected by treatment, $F_1(t)$ still gives useful descriptive information for summarizing the burden of the occurrence of the cause of interest in each of the treatment groups (15). As another example, suppose that eating vegetables reduces death rate by a certain cancer (by

decreasing $h_1(t;z)$). The failure probability for that cancer ($F_j(t;z)$) might not decrease with increasing consumption if survival due to other causes increases with vegetable consumption.

While there have been many studies on the proportional hazards model of Cox, the theory and application of studies on cause-specific failure probability with covariates under the Cox's model are scarce. Further studies are required, for example, on the variances of the functions, the detailed limitations of the usage and interpretations, and so on. This study will contribute to the appropriate choice of a function for the cause-specific failure probability with covariate effects.

APPENDIX

SAS PROGRAMS

I. The Estimate and Standard Error of Net Probability

```
/*********************************************************************/
/* This program estimates the net probability and standard error in simulated data,    */
/* under the assumption of 20% censored, and 80% of cause of interest.                 */
/*********************************************************************/
     /*** Generation of data, 500 subjects, 20% censored, 80% cause of interest ***/
data sim;
  retain seed 1613218064;
  do i=1 to 500;     /* number of subjects in simulated data */
      pyers=(ranuni(seed)*16);     /* survival time since diagnosis, 0-16 years */
       if ranuni(seed)<0.2 then do;     /* 20% censored */
       status=1; cause=.;
       end;
       else if ranuni(seed)>=0.2 then do;     /* 80% nonsensored*/
       status=0 ;
        cause=rantbl(seed, 0.803 , 0.197); /* 80% cause of interest, 20% others */
       end;
       output;
  end;
run;

     /*** Calculation of frequancy of death by cause of interest by survival time ***/
data single; set sim;
   if cause=1 then int=0;  /* cause of interest */
   else int=1;  /* other causes */
proc sort data=single;
   by pyers status;

data single1; set single;
              by pyers;
   if first.pyers then nc1=0;
   if int=0 then nc1+1;
proc sort data=single1;
   by pyers status descending nc1;
data single2; set single1;
              by pyers;
   if first.pyers ne 1 then nc1=.;
   if nd=0 then nc1=.;
```

```
                /*** Calculation of probability of survival beyond time t ***/
                        /*** by treating other causes as censored ***/
data single3; set single2;
              by pyers;
   retain sur 1;
   nr = 501 - _N_; /* number of risk at time t */
   term = 1 - nc1/nr;
   if nc1 =. then term=1;
   sur = sur*term;
   if nc1=. then sur = .;

    /*** Calculation of net probability for prostate cancer occurrence up to time t***/
   fp_sigl=1-sur;

                          /*** Calculation of standard error ***/
   termvar=nc1/(nr*(nr-nc1));
   if nc1=. then termvar=0;
   retain stermvar 0;
   stermvar=stermvar+termvar;
   Var=(sur**2)*stermvar;
   SE=Var**0.5;
proc print;
    var pyers cause nc1 fp_sigl se;
run;
```

II. The Estimate and Standard Error of Crude Probability

```
/*********************************************************************/
/* This program estimates the crude probability and standard error in prostate        */
/* cancer patients in Hawaii under the assumption of homogeneous in characteristics*/
/*********************************************************************/
/************************Basic data processing begins*********************/
                              /*********************/
                              /* Basic Data Processing */
                              /*********************/
          /* This processing will be also used for the programs II, III, and IV */
libname sascds 'c:\1sas\cds';
data pros;
   set sascds.prot (keep=cc age status cause3 diagdate fdate stage);
   if cc=1;
   if stage=9 then delete;  /*deleting of data which have missing value in variable stage*/
/* Categorizing stage */
```

```
   stage1=0; stage2=0;
   if stage=3 then stage2=1;
   if stage=5 or stage=6 or stage=7 then stage1=1;

/*** Calculation of survival time since diagnosis, pyers ***/
   diagyy=substr (diagdate,3,2)+0;
   diagmm=substr (diagdate,5,2)+0;
   diagdd=substr (diagdate,7,2)+0;
   if diagdd=99 or diagdd=00 then diagdd=15;
   diagdate=mdy(diagmm,diagdd,diagyy);
   fdateyy=substr (fdate,3,2);
   fdatemm=substr (fdate,5,2);
   fdatedd=substr (fdate,7,2);
   if fdatedd=99 or fdatedd=00 then fdatedd=15;
   fdate=mdy(fdatemm,fdatedd,fdateyy);
pyers=(fdate-diagdate)/365;

proc sort data=pros;
    by pyers;
data pros1; set pros;
    if _N_<13 then delete; /* Subjects who have zero survival times are deleted */

     /* Calculation of frequancy of all deaths, deaths by prostate cancer, and deaths by
                                other causes by survival time */
    c1 = 0; c2 =0;
    if cause3=1859 then c1=1;  /* death by prostate cancer */
    if cause3 ne 1859 and status=0 then c2=1;  /* death by other causes */
proc sort data=pros1;
    by pyers status descending c2;

data pros2;
    set pros1; by pyers;
    if first.pyers then do;
        nd=0;
        nc1=0;
        nc2=0;
     end;
     if status=0 then nd+1;  /* number of all deaths at time t */
     if c1=1 then nc1+1;  /* number of deaths by prostate cancer at time t*/
     if c2=1 then nc2+1;  /* number of deaths by other causes at time t*/
proc sort data=pros2;
     by  pyers status descending nd;
```

```
data pros3;
     set pros2; by pyers;
     if first.pyers ne 1 or status=1 then do;
     nd=.;
     nc1=.;
     nc2=.;
     end;
/*************************Basic data processing ends************************/

/*** Probability of survival prior to time t, calsur ***/
data pros4; set pros3;
     retain sur 1;
     nr = 435- _N_;    /* risk set at time t, number of subjects=434 */
     term = 1 - nd/nr;
     if nd=. then term=1;
     sur = sur * term;
     lagsur=lag(sur);
     if _N_=1 then calsur=1;
     else calsur=lagsur;
     if nd=. then calsur=.;

/*** Calculation of cause-specific failure probability ***/
     hf = nc1/nr;    /* hazard function estimate for cause 1 at time t */
     fd= hf*calsur;  /* density function estimate for cause 1 at time t */
     if fd=. then fd=0;
     retain fp 0;
     fp=fp+fd;   /* estimate of crude probability */

                         /*** Calculation of standard error ***/
/*Calculation of summation of variance of probability density function estimate at time t*/
    if nr-nd=0 then stvar=.;  /* If denominator is 0, make stvar missing */
    stvar=nd/(nr*(nr-nd));
    lagstvar=lag(stvar);
    if _N_=1 then cstvar=0;
    else cstvar=lagstvar;
    if cstvar=. then cstvar=0;
    retain csvar 0;
    csvar=csvar+cstvar;
    varfd=(fd**2)*((nr-nc1)/(nc1*nr)+csvar);
    if varfd=. then varfd=0;
    retain sumvarfd 0;
    sumvarfd=sumvarfd+varfd;
```

```
data forvar1; set pros4(keep=pyers nc1 nr fp fd csvar sumvarfd);
   if nc1=. or nc1=0 then delete;

/* Calculation of summation of covariance of probability density function estimate */
                                                                    /* at time t */
data a; set forvar1 (keep=fd);
proc transpose data=a out=b (drop=_name_);
data c; set b;
  array x(98) col1-col98; /* 98, number of distinct time of death by prostate cancer */
  array y(98) col1-col98; drop col1-col98;
  do i=1 to 98; do j=1 to 98;
     z=x(i)*y(j);
  output c1;
  end;
  end;
run;

data d; set c1 (keep= i j z);
  if i<=j then delete;
data f; set d;
   retain r 0;
   r=r+z;

data L1; set f;
   if i-j=1;

data L2; set L1;
   lagr=lag(r);
   if _N_=1 then lagr=0;
   unitc=r-lagr;

data L3; unitcov=0;
data L4; set L2 (keep=unitc); unitcov=unitc;
data last; set L3 L4(drop=unitc);

data forvar2; set forvar1(keep=pyers nr nc1 fp csvar sumvarfd);
data covar; set last;
data forvar3; merge forvar2 covar;

data se; set forvar3;
   covfd=unitcov*(-1/nr+csvar);
   retain sumcovfd 0;
   sumcovfd=sumcovfd+covfd;
```

```
   varfp=sumvarfd +2*sumcovfd;  /* variance of estimate of crude probability */
   sefp=varfp**0.5;  /* standard error of estimate of crude probability */

proc print data=se;
   var pyers unitcov covfd fp varfp sefp;
run;
```

III. The Estimate of Cause-specific Failure Probability with Covariate Effects

```
/*********************************************************************/
/* This program estimates the net probability and the models 1-3 for crude probability  */
/* with covariate effects for prostate cancer patients in Hawaii, age=71, stage=localized */
/*********************************************************************/
                        /*******************************/
                        /* Basic Data Processing in program II */
                        /*******************************/
b1=71; b2=0; b3=0;  /*given values of covariates, b1=age, b2 and b3 are dummy      */
                /* variables for stage. If b2=1, stage=regional.  If b3=1, stage=distant.*/

              /*** Estimation of net probability with covariate effects  ***/
data s_mod; set pros3;
     z1=exp(age*0.0313+stage1*1.4132+stage2*2.0686);
      /* Coefficients 0.0313, 1.4132, 2.0686 are estimated from PHREG, for cause 1. */
     lagz1=lag(z1);
     retain sumz1 11621.72;    /*summation of z1 of all observations */
     if _N_=1 then sumz1=11621.72;
     else sumz1=sumz1-lagz1;
     rsumz1=nc1/sumz1;
     if rsumz1=. then rsumz1=0;
     z1b=exp(b1*0.0313+b2*1.4132+b3*2.0686);
     retain crsumz1 0;
     crsumz1=crsumz1+rsumz1;
     chf1=z1b*crsumz1;
     G1co=exp(-chf1);
     single=1-G1co;

/*** Cause-specific failure probability with covariate effects by Model 1 ***/
data mod1; set pros3;
     z1g=exp(age*0.0717+stage1*0.5599+stage2*1.0753);
                        /* Coefficients are estimated from PHREG, for overall deaths. */
     lagz1g=lag(z1g);
     retain sumz1g 123621.22; /*summation of z1g of all observations*/
```

```
    if _N_=1 then sumz1g=123621.22;
    else sumz1g=sumz1g-lagz1g;
    rsumz1g=nd/sumz1g;
    if rsumz1g=. then rsumz1g=0;
    retain G0 1;
    term=1-rsumz1g;
    G0 = G0 * term;
    z1ga=exp(b1*0.0717+b2*0.5599+b3*1.0753);
    Gco = G0**z1ga;
    lagGco=lag(Gco);
    if _N_=1 then sur=1;
    else sur=lagGco;  /* overall survival at time t-1 for given value of covariates */
    if nd=. then sur=.;
    z1=exp(age*0.0313+stage1*1.4132+stage2*2.0686);
    lagz1= lag(z1);
    retain sumz1 11621.72;  /*summation of z1 of all observations */
    if _N_=1 then sumz1=11621.72;
    else sumz1=sumz1-lagz1;
    rsumz1=nc1/sumz1;
    z1a=exp(b1*0.0313+b2*1.4132+b3*2.0686);
    hf=rsumz1*z1a;  /* hazard function estimate for cause 1 at time t */
    fd=sur*hf;  /* density function estimate for cause 1 at time t */
    if fd=. then fd=0;
    retain model_1 0;
    model_1=model_1+fd;

/*** Cause-specific failure probability with covariate effects by Model 2 ***/
data mod2; set pros3;
    z1=exp(age*0.0313+stage1*1.4132+stage2*2.0686);
    lagz1=lag(z1);
    retain sumz1 11621.72;  /*summation of z1 of all observations*/
    if _N_=1 then sumz1=11621.72;
    else sumz1=sumz1-lagz1;
    rsumz1=nc1/sumz1;
    if rsumz1=. then rsumz1=0;
    retain G01 1;
    term1=1-rsumz1;
    G01=G01*term1;
    z1b=exp(b1*0.0313+b2*1.4132+b3*2.0686);
    G1co=G01**z1b;  /* survival for cause 1 at time t for given value of covariates */
    retain G2 1;
    nr=435-_N_;
    term2=1-nc2/nr;
```

```
    if nc2=. then term2=1;
    G2=G2*term2;  /* survival for cause 2 at time t without effect of covariate */
     sur12=G1co*G2;
     lagsur12=lag(sur12);
     if _N_=1 then calsur12=1;
     else calsur12=lagsur12;
     if nd=. then calsur12=.;
     hf=rsumz1*z1b;  /* hazard function estimate for cause 1 at time t */
     fd=calsur12*hf;  /* density function estimate for cause 1 at time t */
     if fd=. then fd=0;
     retain model_2 0;
     model_2=model_2+fd;

/*** Cause-specific failure probability with covariate effects by Model 3 ***/
data mod3; set pros3;
    z1=exp(age*0.0313+stage1*1.4132+stage2*2.0686);
    lagz1=lag(z1);
    retain sumz1 11621.72; /*summation of z1 of all observations*/
    if _N_=1 then sumz1=11621.72;
    else sumz1=sumz1-lagz1;
    rsumz1=nc1/sumz1;
    if rsumz1=. then rsumz1=0;
    retain G01 1;
    term1=1-rsumz1;
    G01=G01*term1;
    z1b=exp(b1*0.0313+b2*1.4132+b3*2.0686);
    G1co=G01**z1b;  /* survival for cause 1 at time t for given value of covariates */
    z2=exp(age*0.0970+stage2*0.4243);
    lagz2=lag(z2);
    retain sumz2 628491.5; /*summation of z2 of all observations*/
    if _N_=1 then sumz2=628491.5;
    else sumz2=sumz2-lagz2;
    rsumz2=nc2/sumz2;
    if rsumz2=. then rsumz2=0;
    retain G02 1;
    term2=1-rsumz2;
    G02=G02*term2;
    z2b=exp(b1*0.0970+b3*0.4243);
    G2co=G02**z2b;  /* survival for cause 2 at time t for given value of covariates */
    sur12=G1co*G2co;
    lagsur12=lag(sur12);
    if _N_=1 then calsur12=1;
    else calsur12=lagsur12;
```

```
    if nd=. then calsur12=.;
    hf=rsumz1*z1b; /* hazard function estimate for cause 1 at time t */
    fd=calsur12*hf; /*density function estimate for cause 1 at time t */
    if fd=. then fd=0;
    retain model_3 0;
    model_3=model_3+fd;
run;

  /*** Merging the results from net probability, model 1, model 2, and model 3 of***/
                                /*** crude probability ***/
data single; set s_mod(keep= pyers single nd nc1);
data model1; set mod1(keep=pyers model_1);
data model2; set mod2(keep=pyers model_2);
data model3; set mod3(keep=pyers model_3);

data mult7100;
  merge single model1 model2 model3; by pyers;
run;

proc print data=mult7100;
  var pyers single model_1 model_2 model_3;
run;
```

IV Graphical Examination of Proportionality Assumption

```
/*********************************************************************/
/* This program examines proportionality assumption for prostate cancer and        */
/*other causes graphically (age=71, stage=localized).                              */
/*********************************************************************/
                        /*********************************/
                        /* Basic Data Processing in program II */
                        /*********************************/
b1=71; b2=0; b3=0;  /*given values of covariates, b1=age, b2 and b3 are
                          dummy variables for stage*/

              /*** Estimation of ln(-ln(Survival) of prostate cancr for ***/
                              /*** given value of covariates */
data pros4; set pros3;
    z1=exp(age*0.0313+stage1*1.4132+stage2*2.0686);
    lagz1=lag(z1);
    retain sumz1 11621.72;
    if _N_=1 then sumz1=11621.72;
    else sumz1=sumz1-lagz1;
```

```
    rsumz1=nc1/sumz1;
    if rsumz1=. then rsumz1=0;
    retain G01 1;
    term1=1-rsumz1;
    G01=G01*term1;
    z1b=exp(b1*0.0313+b2*1.4132+b3*2.0686);
    G1co=G01**z1b;

    ls1=-log(G1co);
    lls1=log(ls1);

            /*** Estimation of ln(-ln(Survival) of other causes for ***/
                        /*** given value of covariates */
    z2=exp(age*0.0970+stage2*0.4243);
    lagz2=lag(z2);
    retain sumz2 628491.5;
    if _N_=1 then sumz2=628491.5;
    else sumz2=sumz2-lagz2;
    rsumz2=nc2/sumz2;
    if rsumz2=. then rsumz2=0;
    retain G02 1;
    term2=1-rsumz2;
    G02=G02*term2;
    z2b=exp(b1*0.0970+b3*0.4243);
    G2co=G02**z2b;

    ls2=-log(G2co);
    lls2=log(ls2);

                        /*** Graphical presentation ***/
data pros5; set pros4(keep=pyers lls1 lls2 nd);
    cause_1=lls1;
    cause_2=lls2;
    if nd=. then delete;

data pros6; set pros5;
    goptions hsize=7.1 in vsize=8.5 in hpos=90 vpos=40;
    symbol1
     value=none
     interpol=join
      line=1
      width=3
      color=black;
```

```
    symbol2
     value=none
     interpol=join
     line=2
     width=3
     color=black;

    axis1 label=(height=1.2 "              ln(-ln(Survival))")
    minor=none;
    axis2 label=(height=1.2 "survival time (years since diagnosis)");

proc gplot data=pros6;
   plot cause_1*pyers
      cause_2*pyers/
      overlay vaxis=axis1 haxis=axis2 frame
      legend;

   footnote1 '        Figure 17. Graphical examination of proportionality ';
   footnote2 '        assumption for prostate cancer and other causes,     ';
   footnote3 '        age 71, stage localized                              ';
   footnote4 '                                                             ';
   footnote5 '                                                             ';
   footnote6 '                                                             ';
   footnote7 '                                                             ';

run;
quit;
```

REFERENCES

1. Prentice RL, Kalbfleisch JD, Peterson AV Jr., Flournoy N, Farewell VT, et al. The analysis of failure times in the presence of competing risks. Biometrics 1978;34:541-554.

2. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York:Wiley,1980.

3. Chiang CL. Introduction to stochastic processes in biostatistics. New York:Wiley,1968.

4. Elandt-Johnson RC, Johnson NL. Survival models and data analysis. New York:Wiley,1980.

5. David HA, Moeschberger ML. The theory of competing risks. London:Griffin,1978.

6. Cornfield J. The estimation of the probability of developing a disease in the presence of competing risks. Am J Public Health 1957;47:601-607.

7. Cox DR, Oakes D. Analysis of survival data. London:Chapman and Hall,1984.

8. Namboodiri K, Suchindran CM. Life table techniques and their applications. Orlando:Academic Press,1987.

9. Karn MN. An inquiry into various death-rates and the comparative influence of certain diseases on the duration of life. Ann Eugenics 1931;4:279-326.

10. Crowley J, Hu M. Covariance analysis of heart transplant survival data. J Am Statist Ass 1977;72(357):27-36.

11. Holt JD. Competing risk analyses with special reference to matched pair experiments. Biometrika 1978;65(1):159-165.

12. Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR et al. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. J Am Statist Ass 1993;88(422):400-409.

13. Cox DR. Regression models and life tables (with discussion) J Royal Statistical Society 1972;B34:187-220.

14. Crowley J, Breslow N. Statistical analysis of survival data. Ann Rev Public Health 1984;5:385-411.

15. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. Biometrics 1990;46:813-826.

16. Farr W. Effect of the extinction of any single disease on the duration of life. Suppl 35th Ann Rep Registrar General 1874;21:38.

17. Todhunter I. A history of the mathematical theory of probability (1865). Reprinted New York: Chelsea Publishing Co.,1949.

18. Makeham WM. On an application of the theory of the composition of decremental forces. J Inst Actuaries. 1874;18:317-322.

19. Fix E, Neyman J. A simple stochastic model of recovery, relapse, death and loss of patients. Human biol 1951;23:205-241.

20. Neyman J. First course in probability and statistics. New York:Henry Holt, 1950.

21. Elveback L. Estimation of survivorship in chronic disease; the actuarial method. J Am Statist Ass 1958;53:420-440.

22. Kimball AW. Models for the estimation of competing risks from grouped data. Biometrics 1969;25:329-337.

23. Berkson J, Elveback L. Competing exponential risks, with particular reference to the study of smoking and lung

cancer. J Am Statist Ass 1960;55:415-428.

24. Pike MC. A note on Kimball's paper "models for the estimation of competing risks from grouped data." Biometrics 1970;26:579-581.

25. Hoel DG. A representation of mortality data by competing risks. Biometrics 1972;28:475-488.

26. Gail MH. A review and critique of some models used in competing risk analysis. Biometrics 1975;31:209-222.

27. Aalen O. Nonparametric inference in connection with multiple decrement models. Scand J Statist 1976;3:15-27.

28. Marshall AW, Olkin I. A Multivariate exponential distribution. J AM Statist Ass 1967;62:30-44.

29. Moeschberger ML, David HA. Life tests under competing causes of failure and the theory of competing risks. Biometrics 1971;27:909-933.

30. Moeschberger ML. Life tests under competing causes of failure. Technometrics 1974;16:39-47.

31. Nadas A. The distribution of the identified minimum of a normal pair determines the distribution of the pair. Technometrics 1971;13:201-202.

32. Seal HL. Studies in the history of probability and statistics, XXXV. Multiple decrements or competing risks. Biometrika 1977;64(3):429-439.

33. Aalen O. Nonparametric estimation of partial transition probabilities in multiple decrement models. Annal Statist 1978;6:534-545.

34. Lawless JF. Statistical models and methods for lifetime data. New York:wiley,1982.

35. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. Annal Statist 1988;16(3):1141-1154.

36. Babu GJ, Rao CR, Rao MB. Nonparametric estimation of specific occurrence/exposure rate in risk and survival analysis. J Am Statist Ass 1992;87:84-89.

37. Huang Y, Wang MC. Estimating the occurrence rate for prevalent survival data in competing risks models. J Am Statist Ass 1995;90(432):1406-1415.

38. Farewell VT. An application of Cox's proportional hazard model to multiple infection data. Appl Statist 1979;28(2);136-143.

39. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika 1978;65:153-158.

40. Kay R. Treatment effects in competing-risks analysis of prostate cancr data. Biometrics 1986;42:203-211.

41. Larson MG, Dinse GE. A mixture model for the regression analysis of competing risks data. Appl Statist 1985;34:201-211.

42. Kuk AYC. A semiparametric mixture model for the analysis of competing risks data. Australian J Statist 1992;34:169-180.

43. Lunn M, McNeil D. Applying Cox regression to competing risks. Biometrics 1995;51:524-532.

44. Davis TP, Lawrance AJ. The likelihood for competing risk survival analysis. Scand J Statist 1989;16:23-28.

45. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Statist Ass 1958;53:457-481.

46. Lee ET. Statistical methods for survival data analysis. Second Edition. New York:Wiley,1992.

47. Miller RG. Survival analysis. New York:Wiley,1981.

48. Greenwood M. The natural duration of cancer. Reports on public health and medical subjects 1926;33, London:Her majesty's stationery office,1-26.

49. Buckley J, James I. Linear regression with censored data. Biometrika 1979;66:429-436.

50. Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. Ann Statist 1981;9:1276-1288.

51. Miller RG. Least squares regression with censored data. Biometrika 1976;63:449-464.

52. Feigl P, Zelen M. Estimation of exponential survival probabilities with concomitant informatiton. Biometrics 1965;21:826-838.

53. Dinse GE, Larson MG. A note on semi-Markov models for partially censored data. Biometrika 1986;73:379-386.

54. Cox DR. Partial likelihood. Biometrika 1975;62(2):269-276.

55. Breslow N. Covariance analysis of censored survival data. Biometrics 1974;30:89-99.

56. Kolonel LN, Yoshizawa CN, Hankin JH. Diet and prostatic cancer; a case-control study in hawaii. Am J Epidemiol

1988;127:999-1012.

57. Le Marchand L, Kolonel LN, Yoshizawa CN. Lifetime occupational physical activity and prostate cancer risk. Am J Epidemiol 1991;133:103-11.

58. Le Marchand L, Kolonel LN, Wilkens LR, Myers BC, Hirohata T. Animal fat consumption and prostate cancer: a prospective study in Hawaii. Epidemiol 1994;5((3):276-282.