AD_____

Award Number: DAMD17-98-1-8256

TITLE: Individual Strategies for Breast Cancer Surveillance Based on Aggregated Familial Information

PRINCIPAL INVESTIGATOR: Andrei Y. Yakovlev, Ph.D.

CONTRACTING ORGANIZATION: University of Utah
Salt Lake City, Utah 84102

REPORT DATE: January 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2001	Annual (1 Jan 00 - 31 Dec 00)

4. TITLE AND SUBTITLE
Individual Strategies for Breast cancer Surveillance Based on Aggregated Familial Information

5. FUNDING NUMBERS
DAMD17-98-1-8256

6. AUTHOR(S)
Andrei Y. Yakovlev, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Utah
Salt Lake City, Utah 84102

E-Mail: yak@hci.utah.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The problem of optimal cancer surveillance is set up as a search for optimal scheduling of medical examinations throughout the lifetime of an individual. Optimal surveillance scheduling strategies allowing for risk variables may be used to further increase the efficacy of breast cancer early detection. To accomplish the general and specific aims of this project, we plan to proceed as follows: (1) develop and implement computer programs for estimating the hazard functions from data on breast cancer incidence; (2) conduct analysis of real data and select significant prognostic variables for a large cohort of women identified through the Utah Population Data Base and the Utah Cancer Registry; (3) construct optimal schedules of breast cancer surveillance and evaluate their potential for enhancing the efficacy of breast cancer detection. This annual report is concerned with estimation of the hazard function and other characteristics of the natural history of breast cancer from epidemiological data.

14. SUBJECT TERMS	15. NUMBER OF PAGES
breast cancer, optimal surveillance, individual strategies	83
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

In Year 2, we were concerned with estimation techniques and analysis of data on breast cancer data from the Utah Population Data Base (UPDB) and the Utah Cancer Registry (UCR). Technical difficulties associated with estimation of the hazard function are described at length in our previous report. All these difficulties have been surmounted and the desired estimates have been obtained from the data amassed in the UPDB and UCR.

1. Statement of Work

This annual report covers the following four tasks formulated in the statement of work.
Task 1. Extraction of breast cancer cohort data from the UPDB and UCR.
Task 2: Development of computer programs for estimation of family history of breast cancer.
Task 3: Development of software for extended hazard regression using linear, quadratic and cubic splines.
Task 4: Evaluation of family history as a predictor of breast cancer on simulated data.
Task 5: Extended hazard regression modeling using familial risk estimates from the breast cancer cohort.
Task 10: Preparation and mailing of the annual report.
Comment: In order to accommodate generally-structured data, we have developed a new methodological approach to the problem of optimal breast cancer surveillance. This is the reason why we began with Tasks 7, 8, and 9 in year 1. This explains why the present report covers Tasks 1, 2, 3, 4, and 5 originally scheduled for year 2.

2. The research carried out to meet the objectives of Tasks 1, 2, 3, 4, 5, 10

2.1. Introduction

In Year 1, we developed several numerical algorithms and software for estimating the hazard function for breast cancer incidence. Allowing for the effects of random censoring and truncation, these procedures have been used for testing covariate effects associated with different indicators of family history.

2.1. Estimation of the hazard rate

Proceeding from preliminary studies of different spline estimation procedures, we chose to model the hazard function via quadratic splines. A quadratic spline with m

knots specifies the hazard to be of the form

$$\lambda_m(t) = \sum_{i=0}^{2} \gamma_{0i} t^k + \sum_{j=1}^{m} \gamma_{j2} (t - \tau_j)_+^2 \qquad (1)$$

where $(x)_+ = \max(x, 0)$. For each birth cohort, we fit splines with knots which are equally spaced in the interior of the interval $[T_{min}, T_{max}]$, where T_{min} is the minimum truncation age in the cohort and T_{max} the maximum follow-up (failure or censoring) time. Restrictions are placed on the coefficients to ensure that $\lambda_m(t)$ remains positive for all t. Thus with m knots the number of parameters is $m + 3$. Models can be fit using maximum likelihood techniques applied to the corresponding conditional likelihood, as discussed in our previous report.

We have developed software designed to compute the spline estimates by maximizing the likelihood function using the algorithm of Powell. We start with one knot and increase the number of knots until the fit is not improved, as determined by the likelihood ratio test at the significance level $\alpha = 0.05$. Three other subcohort estimates of the hazard function were computed for comparison with the spline estimator; an estimator of the life table type, a Gaussian kernel estimate based on the Nelson-Aalen nonparametric estimator, and local likelihood estimators with different kernels (uniform, Epachechnikov, and Gaussian). All the estimators mentioned above are in good agreement with each other when applied to the UPDB data.

Using the computer programs developed in Year 1, the hazard function for cancer incidence has been estimated from left truncated and right censored data on individuals identified through the UPDB and UCR.

Although the estimates become less reliable at increasing age, the hazard function for breast cancer appears to be essentially non-decreasing in all the categories of all familial measures considered. Thus we find no evidence of an "immune fraction" in this analysis. The curves for different levels of risk appear not to merge or cross, indicating that the increased risk to those with a family history does not dissipate after a certain age.

This study is presented at length in the paper by Boucher and Kerber included in Appendix 1.

2.2. Measures of Familial Aggregation as Predictors of Breast Cancer Risk

Several measures of familial disease aggregation have been proposed, but only a few of these are designed to be implemented at the individual level. We have evaluated four of them in the context of breast cancer incidence. After extensive discussions, we came to the conclusion that testing different measures of family history with simulated data was not warranted in view of the fact that such a study would have added little to the results of real data analysis. Therefore, we decided to focus on a more comprehensive analysis of epidemiological data employing a wider spectrum of potential predictors of breast cancer risk.

A population-based cohort consisting of 114,429 women born between 1874 and 1931 and at risk for breast cancer after 1965 was identified by linking the UPDB

and the UCR. Three competing methods were used to obtain predictors of familial aggregation of risk: the number of first degree relatives with breast cancer, the posterior probability of carrying BRCA1 or BRCA2, and the Familial Standardized Incidence Ratio (FSIR), which weights the disease status of relatives based on their degree of relatedness with the proband. Spline regression methods were used to estimate the hazard function, stratified by measures of familial aggregation.

We dichotomized each of our measures of familial risk, with the high risk category representing approximately 8.5% of the data in each case. This was a natural cut point, as it represents the proportion of subjects with one or more first degree relatives with breast cancer. The cutoff for FSIR roughly corresponds to a relative risk of two to family members. The cut points for the posterior probability of BRCA1 and BRCA2 come at points where the posterior probability is rather small, less than 0.0005 in both cases.

Our previous analysis indicated that a highly significant birth-year effect exists in the data, with a women born ten years later having an estimated 40% increased age-specific risk. Birth-year was included as an additional covariate in all regression analyses. The baseline risk was estimated using splines, with the proportional hazards model used for birth-year and familial risk. As with most of the models, we found that two knots were sufficient to provide an optimal fit.

The presence of a first degree relative with breast cancer and the dichotomized FSIR variable each appear to be equally effective at distinguishing high risk subjects, with the high risk category having about double the risk, while the posterior probability of BRCA1 and BRCA2 appear to be less effective.

We performed a more detailed stratified analysis of FSIR. The category boundaries were the approximate 75th, 90th, and 99.9th percentiles of the (adjusted) FSIR distribution. The upper category roughly corresponds to the reported fraction of the general population carrying known breast cancer genes. Bootstrap confidence bands were computed as well as an indicator of the reliability of the estimates.

The estimates of the age-specific hazard and percentile-based bootstrap confidence intervals are presented in Figure 1. The bootstrap confidence intervals are based on 100 bootstrap samples, except for the $<$ 75th percentile category, which is based on 20 bootstrap samples, because of the extensive time it took to fit the models to the large datasets.

We incorporated the posterior probabilities of BRCA1 and BRCA2 and their logarithms, as well as $\log\log FSIR$ as continuous variables in separate analyses, using a proportional hazards model with birth-year as an additional covariate. The best result (in terms of statistical significance) was obtained by including the $\log\log FSIR$, where we get a likelihood ratio $\chi^2_1 = 316.72$ ($p < 0.00001$).

We also considered the indicator variable NFIRST for presence/absence of a first degree relative, in a proportional hazards model. The behavior of the hazard function across different strata shows that the proportional hazards assumption is not grossly violated. The variable NFIRST was highly significant (likelihood ratio $\chi^2_1 = 185.6$, $p < 0.0001$). Addition of a second indicator variable for two or more first degree relatives with breast cancer did not improve the likelihood significantly.

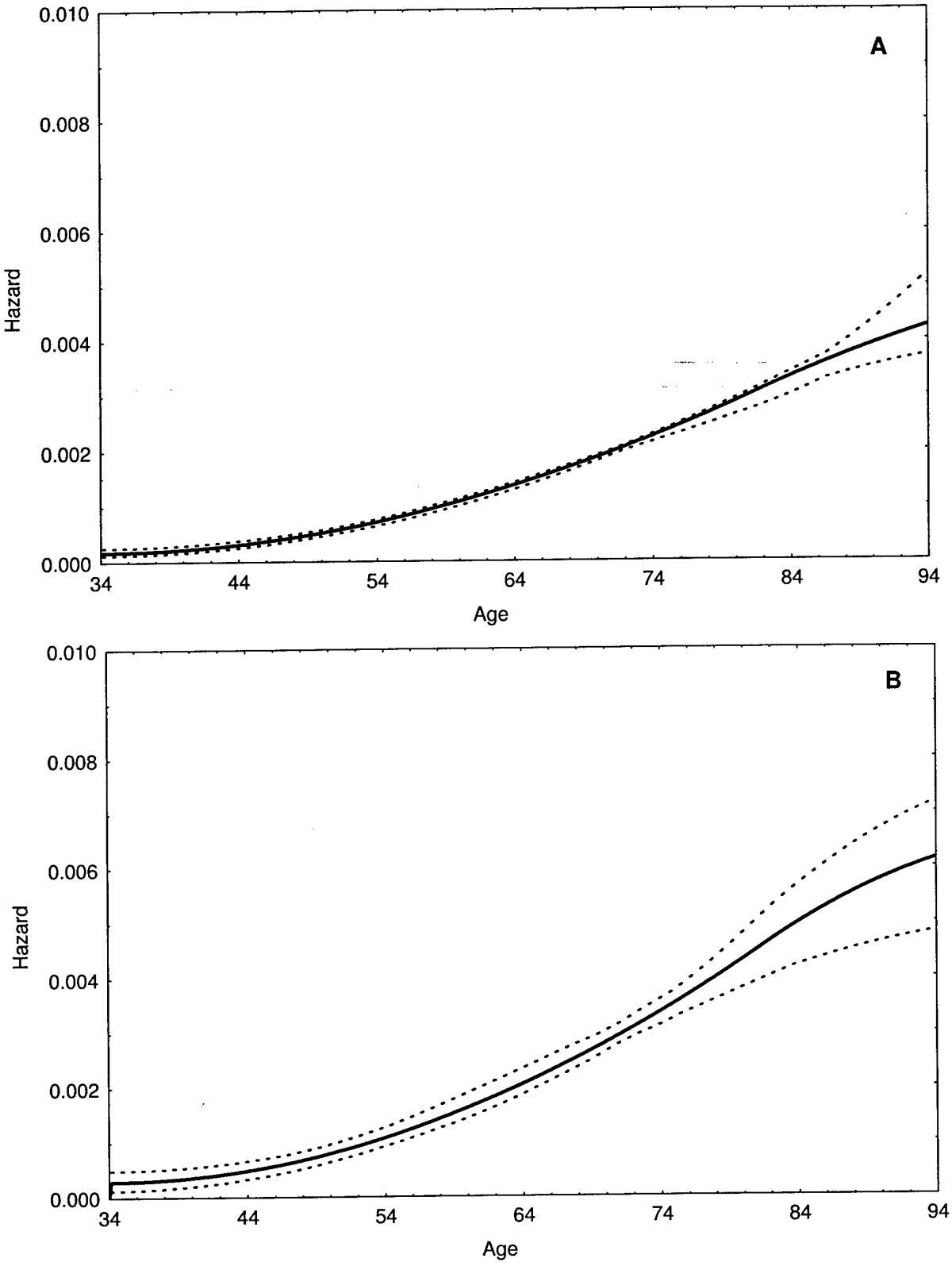
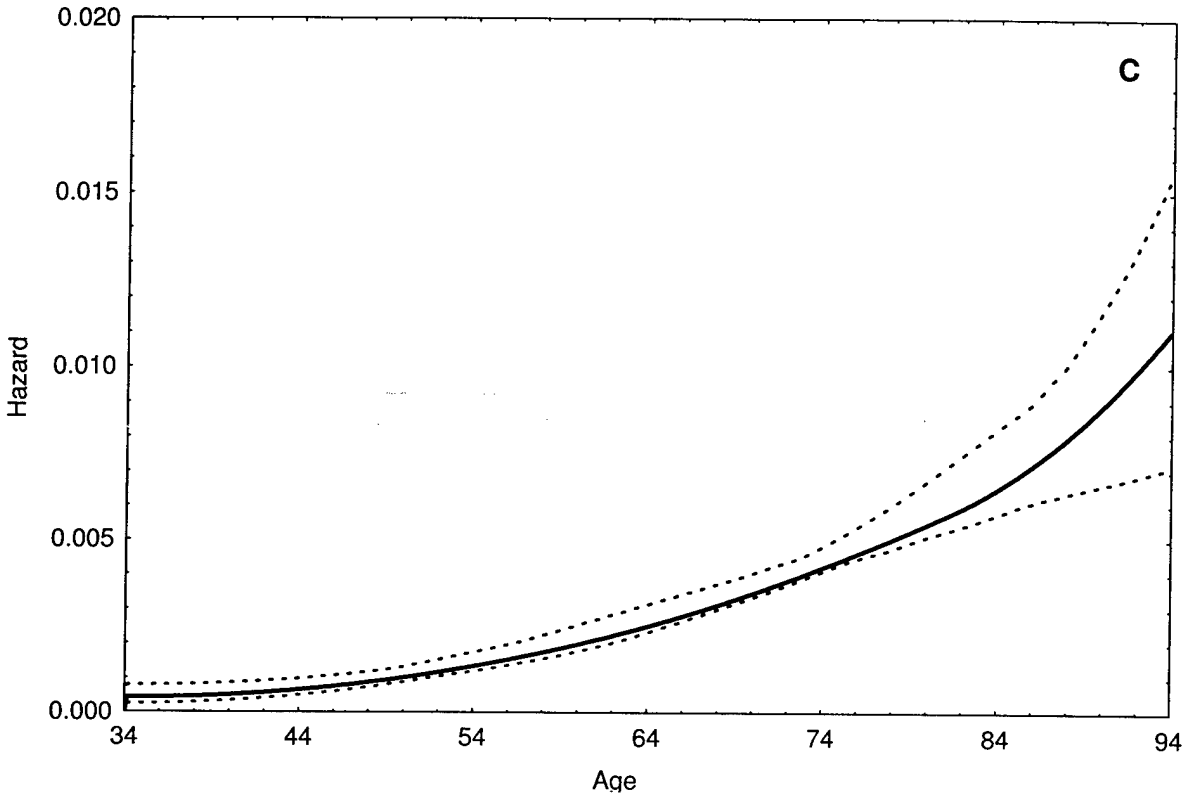


Figure 1. Stratified spline-based estimates and 95% bootstrap confidence bands for the age-specific hazard function for breast cancer. The categories are percentiles 0-75 (A), 75-90(B), 90-99.9 (C), and 99.9-100 (D) of the adjusted FSIR distribution. The scales are different, for better resolution.



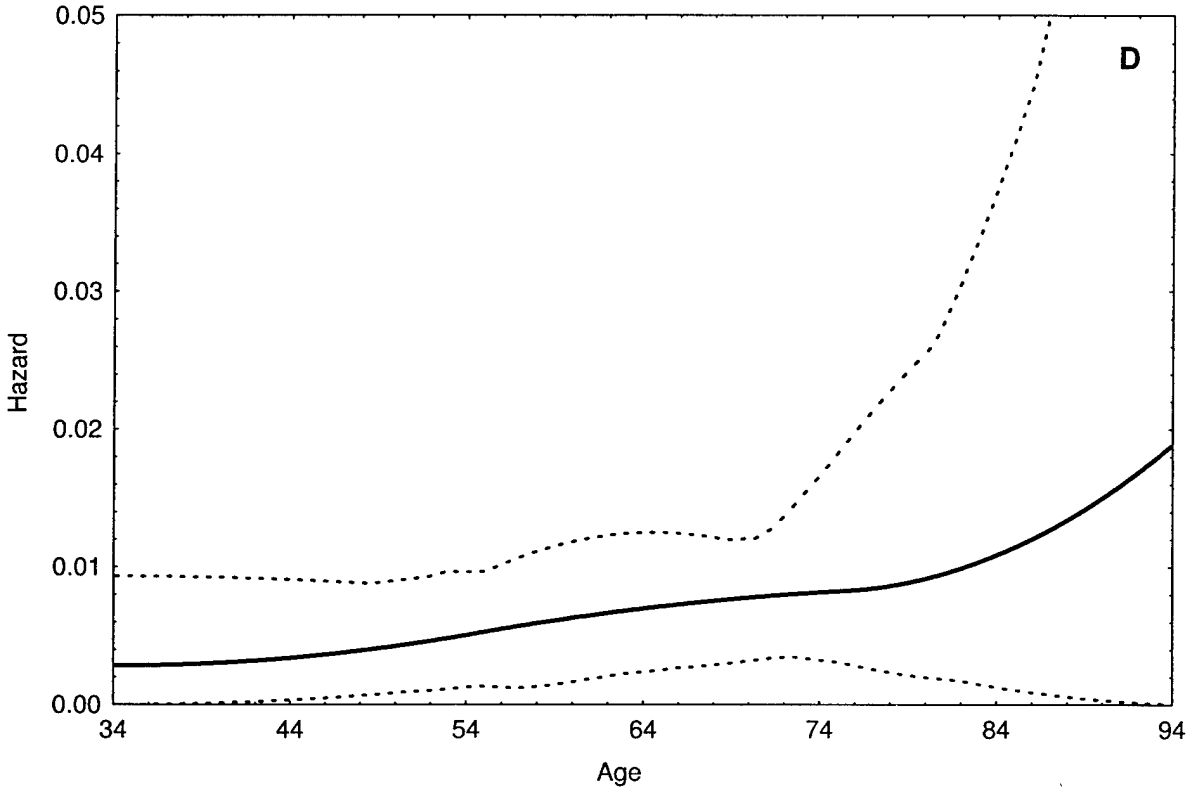


Figure 1 (continued).

More technical details on this study are given in the paper by Boucher and Kerber included in Appendix 2.

2.3. Modeling cancer detection

Let T be the age at tumor onset, and W the time of spontaneous tumor detection measured from the onset of disease. Introduce the random variable (r.v.) S to represent tumor size at spontaneous detection. Then $S = f(W)$, where $f : [0, \infty) \to [1, \infty)$ is a deterministic function describing the law of tumor growth. It is assumed that

(1) random variables T and W are absolutely continuous and independent;

(2) function f is differentiable and $f' > 0$;

(3) the rate of spontaneous tumor detection is proportional to the current tumor size with coefficient $\alpha > 0$.

We observe sample values of the random vector $Y := (T + W, S)$ which components are interpreted as age and tumor size at spontaneous detection, respectively. We look at Y as a transformation of the random vector $X := (T, W)$, $Y = \varphi(X)$, where $\varphi(t, w) = (t + w, f(w))$, $t, w \geq 0$. Observe that components of X are independent random variables. The inverse function $\psi = \varphi^{-1} : A \to \mathbf{R}^2_+$, where $A := \{(u, v) \in \mathbf{R}^2_+ : 1 \leq v \leq f(u)\}$, is given by $\psi(u, v) = (u - g(v), g(v))$, with $g := f^{-1}$. Note that the Jacobian of ψ is g'. Then for the probability density function (p.d.f.) of Y we have assuming that $(u, v) \in A$:

$$p_Y(u, v) = p_X(\psi(u, v))g'(v) = p_T(u - g(v))p_W(g(v))g'(v)$$

$$= p_T(u - g(v))p_S(v).$$

In the particular case of exponential tumor growth with rate $\lambda > 0$ ($f(w) = e^{\lambda w}$) we obtain

$$p_Y(u, v) = \frac{\alpha}{\lambda} e^{-\frac{\alpha}{\lambda}(v-1)} p_T(u - \frac{\ln v}{\lambda}) , \quad u \geq 0, \ 1 \leq v \leq e^{\lambda u} . \tag{2}$$

Thus, the distribution of random vector Y is absolutely continuous but the support of Y depends on the unknown parameter λ. As far as the asymptotic likelihood inference is concerned, the usual regularity conditions are not met for the distribution $p_Y(u, v)$. However, experience with similar parametric settings suggests that the estimation efficiency for the parameter λ may be expected to be even higher than in the regular case although asymptotic normality may fail.

Let $\{(u_i, v_i) : 1 \leq i \leq n\}$ be sample data on age and tumor size at detection. The structure of the joint distribution (2) suggests the following maximum likelihood procedure for estimation of the parameters θ and λ:

(1) Denote $\theta = \alpha/\lambda$ in formula (2), and find the maximum likelihood estimate, $\hat{\theta}$, of the parameter θ using only the tumor size data $\{v_i : 1 \leq i \leq n\}$. It follows (see below) that the sample $\{v_i\}$ is drawn from an exponential distribution with parameter θ, and consequently

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} v_i - 1}.$$

(2) Maximize the function

$$L(\lambda) = \prod_{i=1}^{n} p_T(u_i - \frac{\ln v_i}{\lambda}), \quad u_i > 0, v_i \geq 1,$$

or its logarithm, to find the estimate of λ denoted by $\hat{\lambda}$.
(3) The maximum likelihood estimate of α is given by $\hat{\alpha} = \hat{\theta}\hat{\lambda}$.

The above procedure does the same job as maximizing the likelihood function based on the joint distribution (2). To show this, let the joint density of the random variables U and V be of the form

$$p(u, v; \lambda, \theta) = f(u - \frac{\varphi(v)}{\lambda})g(v; \theta),$$

where $u > 0, v \geq 1, \lambda > 0$. It is assumed that $f(t) > 0$ for $t > 0$, $f(t) = 0$ for $t \leq 0$, $g(x) > 0$ for $x \geq 1$, and $\varphi : [1, \infty) \to (0, \infty)$ is a measurable function. Suppose that there exists a unique maximizer $(\hat{\lambda}, \hat{\theta})$, $\hat{\lambda} > 0$, for the likelihood function

$$L(\lambda, \theta) = \prod_{i=1}^{n} p(u_i, v_i; \lambda, \theta).$$

Then $u_i - \varphi(v_i)/\hat{\lambda} > 0$ for all i, whence

$$\hat{\lambda} > \max_{1 \leq i \leq n} \frac{\varphi(v_i)}{u_i} > 0.$$

Given $\lambda > 0$, it is clear that $\hat{\lambda}$ and $\hat{\theta}$ are unique maximizers for the functions

$$L_1(\lambda) = \prod_{i=1}^{n} f(u_i - \frac{\varphi(v_i)}{\lambda}), \qquad \text{and} \qquad L_2(\theta) = \prod_{i=1}^{n} g(v_i; \theta),$$

respectively. Conversely, if $\hat{\lambda} > 0$ and $\hat{\theta}$ are unique maximizers for these functions, the pair $(\hat{\lambda}, \hat{\theta})$ is a unique maximizer for the likelihood function $L(\lambda, \theta)$. Finally, observe that $g(v; \theta)$ is the marginal density of the random variable V. Indeed, we have

$$\int_0^\infty f(u - \frac{\varphi(v)}{\lambda})g(v; \theta)du = g(v; \theta) \int_0^\infty f(t)dt = g(v; \theta).$$

The performance of the above described estimation procedure was studied by computer simulations. A total of 50 pseudo-random samples of (u_i, v_i) were generated from the joint distribution (2); each sample contained $n = 100$ realizations of the random vector (U, V). We used the composition method to simulate samples of pairs (u_i, v_i). In accordance with this method, we first draw v_i from the marginal distribution of the random variable V, and then generate u_i from the distribution of U conditional on $V = v_i$. The p.d.f. $p_T(x)$ was specified by the Moolgavkar-Venzon-Knudson model of carcinogenesis with the survival function given by the following formula:

$$\bar{G}_T(t) := Pr(T > t) = \left[\frac{(A + B)e^{At}}{B + Ae^{(A+B)t}}\right]^{\delta}, \quad t \geq 0,$$

where $A,\ B,\ \delta > 0$ are identifiable parameters of the model. We used the following values of model parameters: $\alpha = 2.3 \times 10^{-10}, \lambda = 6.9, A = 10^{-4}, B = 0.1821, \delta = 0.0364$.

Simulation Experiment 1. In this experiment, we kept the parameters A, B, and δ at their true values and applied the estimation procedure to simulated data in order to obtain estimates of the parameters λ and α. In this case, the likelihood function can be maximized by a unidimensional search for λ with a fixed value of θ. The estimates of λ and α resulted from each of the 50 samples were summarized by calculating their sample means $\bar{\lambda}$ and $\bar{\alpha}$, as well as the corresponding standard errors (of the sample mean) denoted by $\sigma_{\bar{\lambda}}$ and $\sigma_{\bar{\alpha}}$, respectively. We obtained the following numerical values: $\bar{\lambda} = 7.45$, $\sigma_{\bar{\lambda}} = 0.9$, $\bar{\alpha} = 2.53 \times 10^{-10}$, $\sigma_{\bar{\alpha}} = 0.34 \times 10^{-10}$. These results testify that, given the parameters A, B and δ are known, the estimation procedure performs well when applied to finite samples.

Simulation Experiment 2. Proceeding from the same true parameter values, the estimation procedure was applied to simulated data to obtain estimates of all the parameters incorporated into the model. Since there were three additional parameters to be estimated from simulated data, the size of each sample was increased up to 1000. The results were summarized in just the same way as in Experiment 1 to give: $\bar{\lambda} = 9.4$, $\sigma_{\bar{\lambda}} = 0.9$, $\bar{\alpha} = 3.1 \times 10^{-10}$, $\sigma_{\bar{\alpha}} = 3.1 \times 10^{-11}$, $\bar{A} = 9.5 \times 10^{-4}$, $\sigma_{\bar{A}} = 3.6 \times 10^{-4}$, $\bar{B} = 0.1407$, $\sigma_{\bar{B}} = 0.0599$, $\bar{\delta} = 0.0507$, $\sigma_{\bar{\delta}} = 0.006$.

Simulation Experiment 3. The estimation procedure was applied to a single sample of size 50,000 generated from the joint distribution (2). The estimated parameter values were: $\hat{\lambda} = 6.7, \hat{\alpha} = 2.24 \times 10^{-10}, \hat{A} = 5.1 \times 10^{4}, \hat{B} = 0.1390, \hat{\delta} = 0.0475$.

The above simulation experiments show that estimation of the whole set of model parameters is feasible given the model is adequate for the processes under study, but obtaining unbiased estimates would require large sample sizes.

Suppose now that the process of tumor growth is described by the exponential law $f(w) = e^{\lambda w},\ w \geq 0$, with a *random* growth rate λ. We also assume that the random parameter $\theta := \alpha/\lambda$ is gamma distributed with parameters a and b. Compounding (2) with respect to the gamma distribution of the parameter θ we find the p.d.f. of the resulting randomized distribution of the vector Y :

$$p(u,v) = \frac{b^a}{\Gamma(a)} \int_0^{\alpha u/\ln v} t^a e^{-(b+v-1)t} p_T(u - \frac{\ln v}{\alpha} t) dt\ , \quad u \geq 0,\ v \geq 1\ .$$

Setting $s := u - (\ln v/\alpha)t$ we rewrite the last formula in an equivalent form

$$p(u,v) = \frac{b^a}{\Gamma(a)} \left(\frac{\alpha}{\ln v}\right)^{a+1} \int_0^u (u-s)^a \exp\{-\frac{\alpha}{\ln v}(b+v-1)(u-s)\} p_T(s) ds, \qquad (3)$$

for $u \geq 0$, $v \geq 1$. Alternatively, we may assume that it is the parameter $1/\lambda$ that is gamma distributed with parameters a and b. Should this be the case, we have

$$p(u,v) = \frac{\alpha b^a}{\Gamma(a)} \int_0^{u/\ln v} t^a \exp\{-[b + \alpha(v-1)]t\} p_T(u - t\ln v) dt$$

$$= \frac{\alpha b^a}{(\ln v)^{a+1}\Gamma(a)} \int_0^u (u-s)^a \exp\{-\frac{b+\alpha(v-1)}{\ln v}(u-s)\} p_T(s) ds, \qquad (4)$$

for $u \geq 0$, $v \geq 1$.

Once the density p_T of the age at tumor onset T is specified within a certain parametric family, equations (3) or (4) allow us to compute p.d.f. of the joint distribution of age and tumor size at detection. Observe that in this randomized version the support $[0, \infty) \times [1, \infty)$ of the distribution of random vector Y is parameter free. The maximum likelihood parametric inference based on the joint p.d.f. $p(u, v)$ accommodates censored observations under the usual independent censorship model.

2.4. Future Plans

Formulas (2) and (3) will be used to estimate the natural history of breast cancer from the UPDB data. This will allow us to find a parametric estimate of the p.d.f. $p_{T+W}(t)$, which is necessary for designing optimal schedules of breast cancer screening allowing for information on family history.

3. Key Research Accomplishments

Our key accomplishments in Year 2 can be summarized briefly as follows:

- We have used computer programs developed in Year 1 to estimate the hazard function from data on breast cancer amassed in the UPDB and UCR. These nonparametric estimates are in good agreement with predictions based on the proposed mechanistic model of cancer development and detection.

- We have tested several aggregated measures of family history as predictors of breast cancer risk. This study points the way for data stratification required for construction of individualized strategies of breast cancer surveillance.

- We have derived the joint distribution of tumor size and age at detection and its randomized counterpart which are necessary for estimation of the natural history of the disease. Simulation experiments have been conducted to evaluate how well unknown parameters incorporated into the distribution can be estimated by the maximum likelihood method from available bivariate data on tumor size and age at diagnosis of breast cancer.

4. Reportable Outcomes

4.1. New Publications

1. Yakovlev, A.Y., Tsodikov, A.D., and Hanin, L.G. Optimal schedules of breast cancer surveillance, Abstract, Era of Hope Meeting, Atlanta, June 2000.
2. Boucher, K.M. and Kerber, R.A. The shape of the hazard function for cancer incidence, Abstract, Era of Hope Meeting, Atlanta, June 2000.
3. Boucher, K.M. and Kerber, R.A. The Shape of the Hazard Function for Cancer Incidence, *Mathematical and Computer Modelling*, to appear.

4. Boucher, K.M. and Kerber, R.A. Measures of Familial Aggregation as Predictors of Breast Cancer Risk, *Journal of Epidemiology and Biostatistics*, under revision.

4.2. Awards

1. Grant 1 U01 CA88177-01, NIH/NCI, Mechanistic Modeling of Breast Cancer Surveillance, RFA "Cancer Intervention and Surveillance Network (CISNET)", P.I.: Yakovlev, A.Y., 09/01/00 - 08/31/04, total costs: $ 537,653.

5. Conclusions

The results of data analysis are consistent with an increasing hazard for breast cancer incidence through age 85 or 90. The hazard function appears to be higher for more recent birth cohorts. The shape of the hazard function appears to be consistent with a two-stage model for spontaneous carcinogenesis in which the initiation rate is constant or increasing.

We have applied several methods of measuring familial aggregation at the individual level to breast cancer data. All prove to be significant predictors of individual risk. Judging by the difference in risk estimates, as well as the likelihood ratio test, presence of a first degree relative and FSIR appear to be better indicators of increased risk than the posterior probability of BRCA1 or BRCA2. Judging solely by the likelihood ratio test, one would prefer FSIR. The latter indicator may be thought of as an extension of the cruder number of first degree relatives with breast cancer, adjusting for the level of relatedness and expected disease. It is therefore not surprising to find that it performs better.

Marginal distributions of tumor size and age at detection as well as associated estimation problems were discussed in our previous report. Now we have derived the joint distribution of these two random variables and its randomized counterpart. Generally speaking, explicit formulas for the marginal distributions of tumor size and age of an individual at detection are not sufficient to utilize completely the information contained in the corresponding sample observations for estimation of the natural history of the disease; one needs to know their joint distribution in order to develop pertinent methods for the maximum likelihood statistical inference.

6. So what?

1. As evidenced by the results of data analysis, the shape of the hazard function for breast cancer incidence is consistent with predictions based on the proposed mechanistic model of cancer development and detection.
2. We now know how the data should be stratified with respect to aggregated characteristics of family history in order to construct individualized optimal strategies of breast cancer screening.
3. In Year 3, our focus will be on the development of methods for parametric estimation of the natural history of breast cancer based on formulas (3) and (4) from the

UPDB data stratified with respect to individual information on family history. This study will produce estimates to be used for designing optimal schedules of breast cancer screening.

Appendix 1

The Shape of the Hazard Function for Cancer Incidence

Kenneth M. Boucher and Richard A. Kerber

Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, 2000 East North Campus Drive, Salt Lake City, Utah 84112

Running title: **Hazard Function for Incidence**

Corresponding author:
Kenneth M. Boucher, Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, 2000 Circle of Hope Dr., Salt Lake City, UT 84112, U.S.A.
Phone: 801-585-9544, FAX: 801-585-5357
e-mail: ken.boucher@hci.utah.edu
7/00

ABSTRACT

A population-based cohort consisting of 126,141 men and 122,208 women born between 1874 and 1931 and at risk for breast or colorectal cancer after 1965 was identified by linking the Utah Population Data Base and the Utah Cancer Registry. The hazard function for cancer incidence is estimated from left truncated and right censored data based on the the conditional likelihood. Four estimation procedures based on the conditional likelihood are used to estimate the age-specific hazard function from the data; these were the life-table method, a kernel method based on the Nelson Aalen estimator, a spline estimate, and a proportional hazards estimate based on splines with birth year as sole covariate.

The results are consistent with an increasing hazard for both breast and colorectal cancer through age 85 or 90. After age 85 or 90 the hazard function for female breast and colorectal cancer may reach a plateau or decrease, although the hazard function for male colorectal cancer appears to continue to rise through age 105. The hazard function for both breast and colorectal cancer appears to be higher for more recent birth cohorts, with a more pronounced birth-cohort effect for breast cancer than for colorectal cancer. The age specific hazard for colorectal cancer appears to be higher for men than for women. The shape of the hazard function for both breast and colorectal cancer appear to be consistent with a two-stage model for spontaneous carcinogenesis in which the initiation rate is constant or increasing. Inheritance of initiated cells appears to play a minor role.

KEYWORDS: hazard function, truncation, survival analysis, breast cancer, colorectal cancer

1. Introduction

The shape of the hazard function may lead to insights into the biology of carcinogenesis which may not be easily discernable from a study of the survival function alone. For example, it is typical in the analysis of tumor recurence data to find a hazard function that is bimodal or unimodal, and that tends to zero as time tends to infinity [1]. The modes of the hazard may be interpreted biologically as arising from two different types of failure, one that tends to occur earlier and one that tends to occur later. The decrease in the hazard function to zero may lead one to conclude that there is a non-zero cured fraction. In fact, if we let $\lambda(t)$ denote the hazard function, and p the probability of cure, it follows from the formula

$$p = \lim_{t\to\infty} \exp\left\{-\int_0^t \lambda(u)du\right\},$$

that there are individuals who have been "cured" in the population exactly when the hazard function has finite integral. In particular, $\lim_{t\to\infty} \lambda(t) = 0$, provided the limit exists.

If the hazard function under study is from disease incidence, the "cured fraction" must be re-interpreted as the fraction of the population that is "immune" to the disease. If the cumulative hazard appears to be bounded, for example, one should expect the existence of a non-zero immune fraction. More generally, a large degree of heterogeneity in disease susceptibility may lead to a population hazard function with one or more well-defined maxima. The maxima may correspond to discrete subpopulations with different genetic predisposition to disease. A maximum may also result from a continuous frailty, as the surviving population at higher ages may be overrepresented by individuals with lower risk [2].

Both breast and colorectal cancer are syndromes in which an inherited susceptibility has been shown to play a role. Inherited mutations in p53, BRCA1, BRCA2, the ataxia-telangiectasia gene (AT), HRAS, and the androgen receptor gene (AR) have been shown to play a role in breast cancer susceptibility [3]. About 56% of carriers of the mutation BRCA1 or BRCA2 will get breast cancer by the age of 70 years [4]. BRCA1 has an estimated allele frequency of between 0.0002 and 0.001 (95% CI) [5], and accounts for about 3% of diagnosed breast cancer [6]. The allele frequency of mutations in BRCA2 is estimated at 0.00022 [7]. Germline mutations

in p53 and AR are extremely rare, and mutations in the HRAS1 minisatellite locus which confer increased risk of breast cancer are also rare, having an estimated population frequency of 6% [3]. In a study of 100 Finnish breast cancer families analyzed by protein truncation tests and direct sequencing, Vehmanen et al. [8] found that only 21% of breast cancer families were accounted for by mutations of BRCA1 and BRCA2, providing indirect evidence for the existence of other, undiscovered breast cancer genes.

Indirect evidence also exists for the existence of additional colorectal cancer genes. Inherited mutations in polyposis coli (APC) gene and the hereditary non-polyposis colon cancer syndrome (HNPCC) genes hMSH2, and hMLH1 have been shown to play a role in colon cancer susceptibility [3]. After segregation analysis of 203 pedigrees, Houlston et al. [9] concluded that dominant colorectal cancer genes with a frequency of 0.006 account for an estimated 81% of colorectal cancers in patients under 35, 59% in patients between 35 and 49, decreasing to 16% in patients over 65. The I1307K mutation of the APC gene, found in Ashkenazi Jews, confers an estimated relative risk of 1.7 for colorectal cancer (95% CI 1.01-2.87) [10]. APC and HNPCC are rare, and contribute to a small percentage of colorectal cancer cases [3].

Additional insight can be gleaned from the hazard function for cancer incidence in the framework of a mechanistic model of carcinogenesis. The most widely accepted model is the Moolgavkar-Venzon-Knudson two-stage clonal expansion model [11,12]. The Moolgavkar-Venzon-Knudson model has the following assumptions:
(A) Normal, susceptible target cells are initiated according to a (nonhomogeneous) Poisson process with intensity $\nu(t)$.
(B) The expansion of the colony of initiated cells and malignant transformation is specified by a stochastic birth-death-migration process with the division, death (or differentiation) and transformation. Premalignant cells either divide into two premalignant cells with rate $\alpha(t)$, die with rate $\beta(t)$, or divide asymmetrically into one premalignant cell and one malignant cell with rate $\mu(t)$.

It has been shown that the hazard function for the Moolgavkar-Venzon-Knudson model with constant parameters increases monotonically and approaches an asymptote [13]. An asymptotic value for the hazard is also reached for the Moolgavkar-

Venzon-Knudson model with piecewise constant parameters, and in that case the value of the asymptote depends only on the value of the coefficients in the unbounded interval [13,14].

Expressions for the survivor function were first obtained by Moolgavkar and Luebeck [13]. A simple explicit formula for the survivor function $S(t)$ for the Moolgavkar-Venzon-Knudson model with constant parameters was obtained by Kopp-Schneider et al. [15] and Zheng [16]:

$$S(t) = \left[\frac{2ce^{0.5(-\alpha+\beta+\mu-c)t}}{(-\alpha+\beta+\mu+c)+(\alpha-\beta-\mu+c)e^{-ct}} \right]^{\nu/\alpha} \tag{1}$$

where $c = \sqrt{(\alpha+\beta+\mu)^2 - 4\alpha\beta}$. Zheng also presented an expression for the probability generating function for the number of malignant cells given a single malignant cell at time $t = 0$, allowing an expression for the promotion time distribution

$$F(t) = \frac{(\alpha-\beta-\mu+c)(\alpha-\beta-\mu-c)e^{-ct} + (\alpha-\beta-\mu+c)(-\alpha+\beta+\mu+c)}{2\alpha[(\alpha-\beta-\mu+c)e^{-ct} + (-\alpha+\beta+\mu+c)]} \tag{2}$$

to be given. It is easy to see that $S(t)$ and $F(t)$ above are related by the formula

$$S(t) = \exp\left\{-\nu \int_0^t F(x)dx\right\} \tag{3}$$

which was shown by Hanin and Yakovlev [17] to be valid in a more general setting.

Yakovlev and Tsodikov [18] replace assumption (B) above with the following assumption:

(C) Progenitor cells are transformed into malignant lesions at a random with cumulative distribution function $F(x)$. All progenitor cells are promoted independently of one another.

Assuming $F(0) = 0$, it follows that the process of malignant transformation is also a Poisson process, with integral rate $\Lambda(t) = \int_0^t \nu(u)F(t-u)du$. As in the Moolgavkar-Venzon-Knudson model, the simplest model of spontaneous carcinogenis takes $\nu(t) = \nu$ to be constant, in which case $\Lambda(t) = \nu \int_0^t F(u)du$ and the hazard function for time-to-tumor, given by $\lambda(t) = \nu F(t)$, is nondecreasing. The probability $S(t)$ that there are no malignancies by time t is then given by (3).

This model may easily be modified to handle inherited lesions, via the limiting case where ν is taken to be a delta function at the origin. If $F(t)$ is assumed to be

absolutely continuous, then the integral rate $\Lambda(t)$ is equal to $\nu F(t)$ and the hazard function $\lambda(t) = F'(t) = f(t)$, where $f(t)$ is the density function associated with $F(t)$. We see that the hazard function for spontaneous and inherited lesions are quite likely to have very different shapes.

Even though a thorough study of the hazard function may lead to new insight into the process of carcinogenesis, few if any population-based cohorts have been analyzed to determine the hazard function for cancer incidence. In addition, time-dependent variation in environmental risk factors for cancer may cause estimates from a cross-sectional study to be misleading. In this paper the age specific hazard function for both breast and colorectal cancer incidence are estimated using data from the Utah Cancer Registry and the Utah Population Data Base. We see that the hazard function for both these types of cancer appears to be increasing monotonically, at least through age 85 or 90. In the context of the above mechanistic models of carcinogenesis, we will see that risks for both these cancers at the population level appear to be relatively homogeneous, with negligible inherited component.

2. Methods

2.1 Data

The data for this study was obtained by linking records from the Utah Population Data Base (UPDB) with the Utah Cancer Registry (UCR). The UPDB consists of the genealogical records of more than 1,000,000 individuals who were born, died, or married in Utah, or en route to Utah during the nineteenth and twentieth centuries. Since 1973 the UCR has been reporting to National Cancer Institutes Surveillance Epidemiology and End Results (SEER) program, and is required to maintain very high standards for case reporting and follow-up, and to periodically undergo quality control audits by SEER personnel to assure uniformly high quality and consistency from year to year. The available follow-up information comes either from Utah death certificates, which have been linked to the UPDB genealogical data every year from 1933 through the beginning of 1997, or from linkage of the HCFA beneficiary data to the UPDB. The study population consisted of 126,141 men and 122,208 women recorded in the Utah Population Database, who were born from 1874 to 1931 and for

whom follow-up information is available that places them in Utah during the years of operation of the Utah Cancer Registry (1966-present). Subjects with purported follow-up past age 105 were excluded from the data. There are 5,372 cases of female breast cancer and 5,177 cases of colorectal cancer represented in the data. Analyses were performed on subcohorts based on birth year (1874-1889, 1890-1899, 1900-1909, 1910-1919, and 1920-1931) and gender. For each gender he entire cohort (birth years 1874-1931) was also analyzed as a whole. The total number of subjects and cases of breast and colorectal cancer for each birth subcohort and gender are given in Tables 1 and 2. Male breast cancer was not analyzed.

2.2 Truncation: Nonparametric Estimation

We wish to estimate the age specific hazard function for breast and colorectal cancer from the data described above, taking into account that the data is subject to random truncation: cases which occurred during or before 1965 are not recorded in the dataset. Subject were between the ages of 34 and 86, at the time of truncation. Thus, analysis of the data must take into account not only to the effects of right censoring, but also the effects of left truncation due to delayed entry into the risk set. The topic of random truncation is not mentioned in several authoritative texts such as Kalbfleisch and Prentice [19] and Fleming and Harrington [20], and may be unfamiliar to some readers, and therefore will be discussed in this and the following subsection.

Let the truncation time Y have distribution function $G(y)$ and the failure time (time of cancer diagnosis) X have distribution function $F(x)$. We require that truncation be independent of failure and for simplicity assume no censoring for the present. Observations are conditional on $X > Y$. Let $G^*(y)$ and $F^*(x)$ be the corresponding distribution functions, conditional on $X > Y$. Let $S(x) = 1 - F(x)$ be the survivor function of X. Suppose that we have observations $(Y_1^*, X_1^*), \ldots, (Y_n^*, X_n^*)$ from the conditional distribution. The full likelihood of the observed data is given by

$$L = \prod_{j=1}^{n} \left[dF(X_j) dG(Y_j)/\alpha \right], \tag{4}$$

where $\alpha = \int \int_{y \leq x} dF(x) dG(y)$. A key observation is that if X and Y are independent, then the hazard of X given $X > Y = y$ at $x > y$ is equal to the hazard of X at x

[21,22]. This observations leads to the result, first mentioned by Kaplan and Meier [23], that if the distribution $G(t)$ is allowed to vary freely, the natural generalization of the product limit estimator, given by the formula

$$\hat{S}(t) = \prod_{X_i^* \leq t} \left(1 - \frac{1}{R(X_i^*)}\right), \tag{5}$$

where $R(u) = \#\{Y_i^* < U \leq X_i^*\}$ is the number at risk at U, is the nonparametric maximum likelihood estimator (NPMLE) of the survivor function $S(t)$ of X (see, for example [21,22,24]).

This result extends naturally to the case with random independent censoring [24]. It also easily follows that in the nonparametric setting (again with no censoring), maximizing (4) is equivalent to maximizing the conditional likelihood of $(X_1^*, \ldots, X_n^*)$ given $(Y_1^*, \ldots, Y_n^*)$, which can be written

$$CL = \prod_{i=1}^{n} f(X_i^*)/S(Y_i^*). \tag{6}$$

(see, for example, [23-26]). Maximizing the conditional likelihood also leads to the familiar Nelson-Aalen estimator for the integrated hazard function $H(t)$ of X [24], which is given by

$$\hat{\Lambda}(t) = \sum_{X_i^* \leq t} R(X_i^*)^{-1}. \tag{7}$$

These results can be extended to the case of right censoring [24].

2.3 Truncation: Parametric Models

We consider the situation where X and Y are independent, $F(x)$ is parametrized, while $G(y)$ is allowed to vary freely. In a later subsection $F(x)$ will be come from a quadratic spline model.

The data are independent pairs $(y_1, x_1), \ldots, (y_n, x_n)$ from the joint distribuition (Y, X), conditional on $(Y < X)$. We suppose, for simplicity, that there are no ties among $y_1, y_2, \ldots, y_n$, and suppose X has absolutely continuous distribution function coming from a family $F(x; \vec{z})$ parameterized by a vector $\vec{z}$, with corresponding survival function $S(x; \vec{z}) = 1 - F(x; \vec{z})$ and density $f(x; \vec{z})$. The NPMLE for G should consist of (unknown) point masses $q_1, q_2, \ldots, q_n$ placed at the points $y_1, y_2, \ldots, y_n$.

The logarithm of the complete likelihood (4) can be rewritten

$$\log(L) = \sum_{i=1}^{n}[\log(f(x_i;\vec{z})) + \log(q_i)] - n\log\left[\sum_{j=1}^{n} S(y_i;\vec{z})q_i\right]. \tag{8}$$

If we factor the out the part of the likelihood corresponding to (6), the logarithm is given by

$$\log(CL) = \sum_{i=1}^{n}[\log(f(x_i;\vec{z})) - \log(S(y_i;\vec{z}))]. \tag{9}$$

We now discuss the changes which must be made in when censoring and additional covariates are present. If $\vec{s}$ is a vector of additional covariates, $\lambda(x,\vec{s};\vec{z})$ denotes the hazard associated with $F(x,\vec{s};\vec{z})$ and $\Lambda(x,\vec{s};\vec{z})$ the cumulative hazard, we note that (9) becomes

$$\log(CL) = \sum_{i=1}^{n}[\log(\lambda(x_i,\vec{s}_i;\vec{z})) - (\Lambda(x_i,\vec{s}_i;\vec{z}) - \Lambda(y_i,\vec{s}_i;\vec{z}))]. \tag{10}$$

In the presence of right censoring which is independent of both the failure and truncation times, x_i is replaced in the above formulation by the minimum of the failure and censoring time. The term $f(x,\vec{s};\vec{z})$ in the likelihood is replaced by $f(x,\vec{s};\vec{z})^{\delta}S(x,\vec{s};\vec{z})^{1-\delta}$, where $\delta_i = 1$ if observation i is a failure and $\delta_i = 0$ otherwise, and the conditional likelihood (6) (with $x_i, \vec{s}_i$ and y_i regarded as fixed) becomes

$$CL = \prod_{i=1}^{n}[f(x_i,\vec{s}_i;\vec{z})^{\delta}S(x_i,\vec{s}_i;\vec{z})^{(1-\delta)}]/S(y_i,\vec{s}_i;\vec{z}).$$

In this setting $\log(CL)$ becomes

$$\log(CL) = \sum_{i=1}^{n}[\delta_i\log(\lambda(x_i,\vec{s}_i;\vec{z}_i)) - (\Lambda(x_i,\vec{s}_i;\vec{z}) - \Lambda(y_i,\vec{s}_i;\vec{z}))]. \tag{11}$$

In the subsequent analysis we choose to maximize (11) rather than the full likelihood.

2.4 Spline Models

We choose to model the hazard via quadratic splines as in [27]. A quadratic spline with m knots specifies the hazard to be of the form

$$\lambda_m(t) = \sum_{i=0}^{2}\gamma_{0i}t^k + \sum_{j=1}^{m}\gamma_{j2}(t-\tau_j)_+^2 \tag{12}$$

where $(x)_+ = \max(x,0)$. For each birth cohort, we fit splines with knots which were equally spaced in the interior of the interior $[T_{min}, T_{max}]$, where T_{min} is the minimum

truncation age in the cohort and T_{max} the maximum follow-up (failure or censoring) time. Restrictions were placed on the coefficients to ensure that $\lambda_m(t)$ remained positive for all t. Thus with m knots the number of parameters was $m+3$. Models were fit using maximum likelihood techniques applied to the conditional likelihood, as given by (11).

The hazard function was estimated for breast cancer incidence (women only) and for colorectal cancer incidence (both men and women). The spline estimates were computed by maximizing $\log(CL)$ using the algorithm of Powell [28]. We started with one knot and increased the number of knots until the fit was not improved, as determined by the likelihood ratio test at the significance level $\alpha = 0.05$. Two other subcohort estimates of the hazard function were computed for comparison with the spline estimator; a life table version of (5), and a Gaussian kernel estimate based on the Nelson-Aalen estimator (7).

2.5 Proportional Hazards

It became clear when fitting models to the subcohorts, that there was a birth cohort effect in the data. At the same time, we wished to have estimates of the hazard for the entire age range of 34-100+ years. We therefore fit proportional hazards models with splines $\lambda_m(t)$ for the baseline hazard and a single covariate s representing birth year. The resulting hazard function has the form

$$\lambda_m(t, s; \beta) = \exp(\beta s)\lambda_m(t). \tag{13}$$

The model was again fit using the conditional likelihood of the form

$$\log(CL) = \sum_{i=1}^{n} [\delta_i \log(\lambda_m(x_i, s_i; \beta)) - (\Lambda(x_i, s_i; \beta) - \Lambda(y_i, s_i; \beta))], \tag{14}$$

which is (11) with $\lambda(x, \vec{s}, \vec{z}) = \lambda_m(x_i, s_i; \beta)$.

3. Results

Estimates of the age specific hazard for for female breast cancer are presented in Figure 1 for the 1874-1889, 1890-1899, 1900-1909, 1910-1919, and 1920-1931 birth subcohorts. Age specific hazards for colorectal cancer are presented in Figures 2

and 3, stratified by birth cohort and gender. Each figure presents three estimates of the hazard from the subcohort alone, namely the life table estimate, the kernel estimate based on the Nelson-Aalen estimator and a spline estimate, as well as and one gender-specific estimate from a proportional hazards model with birth year as covariate, fit to data from all birth subcohorts (1874-1931). The covariate is set to the mean birth year of the subcohort. We note that approximately 40 years of follow up are available for any one subcohort, as follow up data are available from approximately 1965-1995.

We found that splines with very few knots appeared to fit the data. In all but one case two knots were sufficient for the spline estimates, as determined by the likelihood ratio test, and in the remaining case (breast cancer, birth years 1874-1889) one knot sufficed. The hazard function for both breast and colorectal cancer appears to increase monotonically, at least until the age of 85 or 90, when the subcohort specific estimates of the hazard estimates for women for both breast and colon cancer appear to flatten or decrease while the estimate for men appears to continue to increase. (In each of the three cases the proportional hazards model provides estimates of the hazard function which increase through all ages.) We also note that in all the proportional hazards models the birth cohort effect was highly significant (p¡ 0.0001). We also see from the subcohort analysis that the proportional hazards assumption appears to be adequate, at least up until the age of 85 or 90, when proportionality may fail for women.

We also note that the colorectal cancer risk estimates are higher for men than for women. For example, the estimated age specific yearly hazard for the 1920-1931 birth cohort at age 70 is approximately .0013 for women, and about .0017 for men, or about 30% higher for men.

The estimated hazard from the proportional hazards models over a 70 year range are presented in Figures 4 - 6. The estimated hazards increase as the birth cohorts become more recent, with coefficient estimates of $\beta = 0.0347$ (year^{-1}) for female breast cancer, $\beta = 0.016$ (year^{-1}) for female colorectal cancer and $\beta = 0.020$ (year^{-1}) for male colorectal cancer. Thus, the additional hazard for more recent birth cohorts appears to be more pronounced for breast cancer than for colorectal cancer.

4. Discussion

As noted in the Introduction, the presence of a large degree of heterogeneity in the risk for a population may lead to a decreasing age specific hazard function. Since we see little or no evidence of a decreasing hazard for either breast or colorectal cancer at least until age 85 or 90, it appears that the risk is relatively homogeneous for both these cancers over this age range. In particular, there appears to be little evidence for a high immune fraction for either breast or colorectal cancer. We should also note that the presence of a monotone increasing hazard over a limited range does not completely rule out heterogeneity. The data is quite consistant with the degree of heterogeneity that might result from known cancer genes, as long as the risk is generally increasing (at least through age 90) in the population as a whole. There is little or no evidence of an inherited component to the risk, as a large inherited component might be expected to provide a local maxima to the hazard rather early in life, certainly prior to age 85.

One may extend the more general two-stage model of carcinogenesis presented in the Introduction to take cell death into account, by adding a Poisson process of cell death which competes with the process of malignant transformation, as suggested by Yakovlev and Polig [29]. This model has been successfully applied to data from radiation induced and chemically induced lesions [30-32]. With the cell death component it becomes less clear that the hazard function should increase monotonically in the case of spontaneous carcinogenesis. In fact, in the simplified case of constant rates ν_1 of initiation and ν_2 of cell death, and arbitrary cumulative distribution function $F(t)$ for time to transformation of intermediate lesions, the hazard function for time to tumor has the form

$$\lambda(t) = \nu_1 \exp(-\nu_2 t) F(t). \tag{15}$$

We note that according to this model the clock for cell death in this model starts at birth. If the constant $\nu_2 > 0$ in (15), then $\lambda(t)$ must decrease exponentially since $F(t)$ approaches one as t approaches infinity. We conjecture that in the present context the cell death component is very small, so that it does not dominate $\lambda(t)$ until after age 85. The higher hazard rate for male colorectal cancer, as well as the continued increase in hazard through age 105, may be attributed to a smaller rate

of cell death. Another possibily is that the cell death should not be measured from birth, but from formation of the initiated cell (as in another variation of the model suggested in [29]).

We have noted in the Results section that proportionality of hazard appears to fail after age 90 for both breast and colorectal cancer in women. This result may be due to sampling variability, or additional bias unique to women at these high ages. We note that there are only 116 female breast cancer cases and 77 female colorectal cancer cases after age 90. They are distributed over a 15 year period, for an average of 7.7 breast cancer and 5.1 colorectal cancer cases per year in this range. In addition, data linkage is more difficult for women, who are more likely to have changed names than men. An additional indication that the lack of proportionality for women may be spurious is that we do not see this apparent lack of proportionality in men.

5. Acknowledgments

This research was supported, in part, by NCI Cancer Center Support Grants 5P30 CA 4201 and 2P30 CA 42014, U.S. Army Medical Research and Materiel Command Grant DAMD17-98-1-8256, and by NIH/NCI grant R29 CA69421. Partial support was provided by the Huntsman Cancer Institute for the Utah Population Data Base. The Utah Cancer Registry was supported by NCI-PC-67000.

In addition, the authors would like to thank Professor Andrei Y. Yakovlev for many helpful discussions.

REFERENCES

1. A.Y. Yakovlev, A.D. Tsodikov, K. Boucher, and R. Kerber, The shape of the hazard function for breast cancer: curability of the disease revisited. *Cancer* **85**, 1789-1798, (1999).

2. O. O. Aalen, Modeling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability* **2**, 951-972, (1992).

3. D. F. Easton, The inherited component of cancer. *British Medical Bulliten* **50**, 527-535, (1994).

4. J. P. Struewing, P. Hartge, S. Wacholder, S. M. Baker, M. Berlin, M. McAdams, M. M. Timmerman, L. C. Brody, and M. A. Tucker, The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N. Engl. J. Med.* **336**, 1401-1408, (1997).

5. D. Ford and D. F.Easton, The genetics of breast and ovarian cancer. *Br. J. Cancer* **72**, 805-812, (1995).

6. B. Newman, H. Mu, L. M. Butler, R. C. Millikan, P. G. Moorman, and M.C. King, Frequency of breast cancer attributable to BRCA1 in a population-based series of American women *JAMA* **279**, 915-921, (1998).

7. T. I . Anderson, Genetic heterogeneity in breast cancer susceptibility. *Acta Oncol.* **35**, 407-410, (1996).

8. P. Vehmanen, L.S. Friedman, H. Eerola, L. Sarantaus, S. Pyrh:.onen, B.A.J. Ponder, T. Muhonen et al., A low proportion of BRCA2 mutations in Finnish breast cancer families. *Am J. Human Genet.* **60**, 1050-1058, (1997).

9. R. S. Houlston, A. Collins, J. Slack and N. E. Morton, Dominant genes for colorectal cancer are not rare. *Ann. Human Genet.* **56**, 99-103, (1992).

10. S. J. Laken, G. M. Petersen, S. B. Gruber, C. Oddoux, H. Ostrer, G. M. Giardiello, et al, Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat. Gent.* **17**, 79-83, (1997).

11. S. H. Moolgavkar and D. J. Venzon, Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. *Math. Biosci.* **47**, 55-77, (1979).

12. S. H. Moolgavkar and A.Knudson, Mutation and cancer: a model for human carcinogenesis. *J. Natl Cancer Institute* **66**, 1037-1052, (1981).

13. S. H. Moolgavkar and E. G. Luebeck, Two-event model for carcinogenesis: biological, mathematical and statistical considerations. *Risk Anal.* **10**, 323-341, (1990).

14. W. F. Heidenreich, E. G. Luebeck and S. H. Moolgavkar, Some properties of the hazard function of the two-mutation clonal expansion model. *Risk Analysis* **17**, 391-399, (1997).

15. A. Kopp-Schneider, C.J. Portier, and C. D. Sherman, The exact formula for tumor incidence in the two-stage model. *Risk Analysis* **14**, 1079-1080, (1994).

16. Q. Zheng, On the exact hazard and survival functions of the MVK stochastic carcinogenesis model. *Risk Anal.* **14**, 1081-1084, (1994).

17. L. G. Hanin and A. Y. Yakovlev, A nonidentifiability aspect of the the two-stage model of carcinogenesis. *Risk Anal.* **16**, 711-715, (1996).

18. A. Yu. Yakovlev and A. D. Tsodikov, *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, Singapore (1996).

19. J. D. Kalbfleisch and R.L. Prentice, *The statistical analysis of failure time data*, Wiley, New York (1980).

20. T.R. Fleming and D.P. Harrington, *Counting Processes and Survival Analysis*, Wiley, New York, (1991).

21. N. Keiding Independent delayed entry. In *Survival Analysis: the State of the Art*, J. P. Klein and P. K. Goel, eds., Kluwer, Boston-Dordrecht-London, 1992, pp. 309-326.

22. M. Woodroofe, Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163-177, (1985).

23. E. L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481, (1958).

24. N. Keiding and R. D. Gill, Random trucation models and Markov processes *Ann. Statist.* **18**, 582-602, (1990).

25. S. Johansen, The product limit as a maximum likelihood estimator. *Scand J. Statist.* **5**, 195-199, (1978).

26. M.-C. Wang, N. P. Jewell, and W.-Y. Tsai, Asymptotic properties of the product limit estimate under random truncation *Ann. Statist.* **14**, 1597-1605, (1986).

27. J. Etezadi-Amoli and A. Ciampi, Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function. *Biometrics* **43**, 181-192, (1987).

28. D.M. Himmelblau, *Applied Nonlinear Programming* McGraw-Hill, Austin, 1972.

29. A. Yakovlev and E. Polig, A diversity of responses displayed by a stochastic model of radiation carcinogenesis allowing for cell death. *Math. Biosci.* **132**, 1-33, (1966).

30. A. Yakovlev, W. Müller, L. Pavlova and E. Polig, Do cells repair precancerous lesions induced by radiation? *Math. Biosci.* **142**, 107-117, (1997).

31. K.M.Boucher and A.Y. Yakovlev, Estimating the probability of initiated cell death before tumor induction. *Proc. Natl. Acad. Sci. USA* **94**, 12776-12779, (1997).

32. K. Boucher L.V. Pavlova, and A. Y. Yakovlev, A model of multiple tumorigenesis allowing for cell death quantitative insight into biological effects of urethane. *Math. Biosci.* **150**, 63-82, (1998).

Legends to figures

Figure 1. Four estimates of the age-specific hazard function for female breast cancer, stratified by birth cohort: a spline estimate (labeled "Spline"), a kernel estimate based on the Nelson-Aalen estimator (labeled "Kernel"), a life table estimate (labeled "Life Table"), and a proportional hazards spline estimate using all strata, with birth year as sole covariate, set at the stratum mean (labeled "Combined Spline").

Figure 2. Four estimates of the age-specific hazard function for female colorectal cancer, stratified by birth cohort: a spline estimate (labeled "Spline"), a kernel estimate based on the Nelson-Aalen estimator (labeled "Kernel"), a life table estimate (labeled "Life Table"), and a proportional hazards spline estimate using all strata, with birth year as sole covariate, set at the stratum mean (labeled "Combined Spline").

Figure 3. Four estimates of the age-specific hazard function for male colorectal cancer, stratified by birth cohort: a spline estimate (labeled "Spline"), a kernel estimate based on the Nelson-Aalen estimator (labeled "Kernel"), and a life table estimate (labeled "Life Table"), and a proportional hazards spline estimate using all strata, with birth year as sole covariate, set at the stratum mean (labeled "Combined Spline").

Figure 4. Comparison of the age-specific hazard function estimates for female breast cancer for various birth cohort strata from a proportional hazards model spline model. Birth year covariate set at the mean value for each stratum: 1884.41 for the 1874-1889 stratum, 1894.90 for the 1890-1899 stratum, 1904.54 for the 1900-1909 statum, 1914.52 for for the 1910-1919 statum, and 1925.24 for the 1920-1931 stratum.

Figure 5. Comparison of the age-specific hazard function estimates for female colorectal cancer for various birth cohort strata from a proportional hazards model spline model. Birth year covariate set at the mean value in each stratum: 1884.41 for the 1874-1889 stratum, 1894.90 for the 1890-1899 stratum, 1904.54 for the 1900-1909 statum, 1914.52 for for the 1910-1919 statum, and 1925.24 for the 1920-1931 stratum.

Figure 6. Comparison of the age-specific hazard function estimates for male colorectal cancer for various birth cohort strata from a proportional hazards model spline model. Birth year covariate set at the mean value in each stratum: 1884.74 for the 1874-1889 stratum, 1895.06 for the 1890-1899 stratum, 1904.74 for the 1900-1909 statum, 1914.57 for for the 1910-1919 statum, and 1925.31 for the 1920-1931 stratum.

Table 1. Number of female subjects and cases of breast and colorectal cancer, stratified by birth year.

Birth Years	Number of Subjects	No. of breast cancer cases	No. of colorectal cancer cases
1874-1889	10,115	145	116
1890-1899	19,352	564	435
1900-1909	27,138	1,258	755
1910-1919	31,162	1,709	752
1920-1931	34,441	1,696	448
Total	122,208	5,372	2,106

Table 2. Number of male subjects and cases of colorectal cancer, stratified by birth year.

Birth Years	Number of Subjects	No. of colorectal cancer cases
1874-1889	6,850	101
1890-1899	16,307	341
1900-1909	27,122	768
1910-1919	34,731	874
1920-1931	41,131	587
Total	126,141	2671

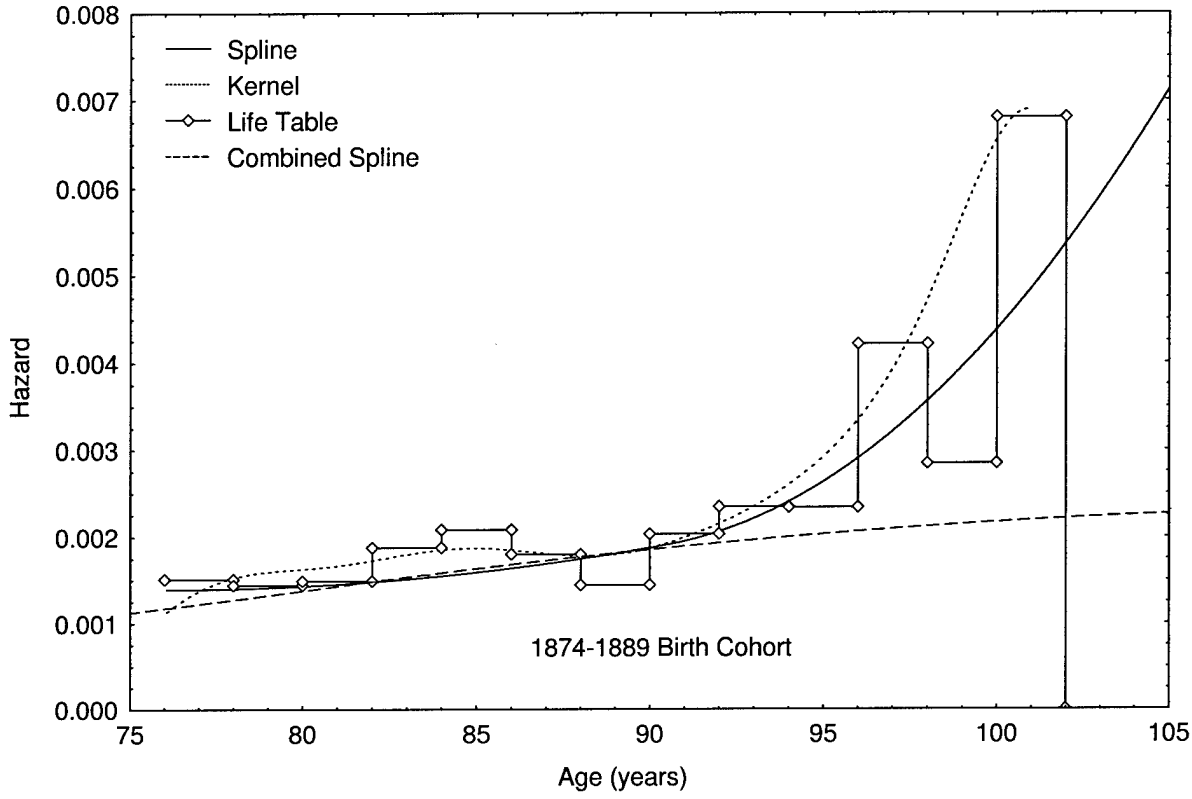

Spline
Kernel
Life Table
Combined Spline
Hazard
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
1874-1889 Birth Cohort
75
80
85
90
95
100
105
Age (years)

FIGURE 1A

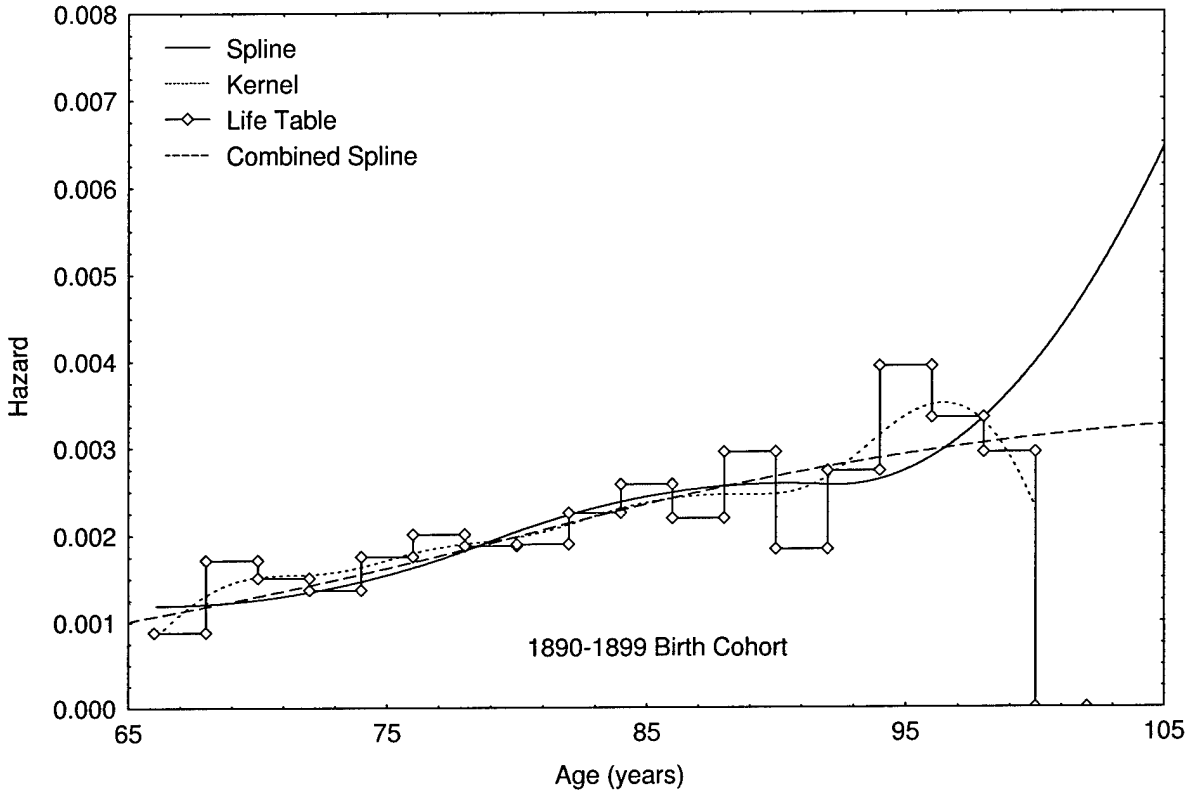

Spline
Kernel
Life Table
Combined Spline
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Hazard
1890-1899 Birth Cohort
65
75
85
95
105
Age (years)

FIGURE 1B

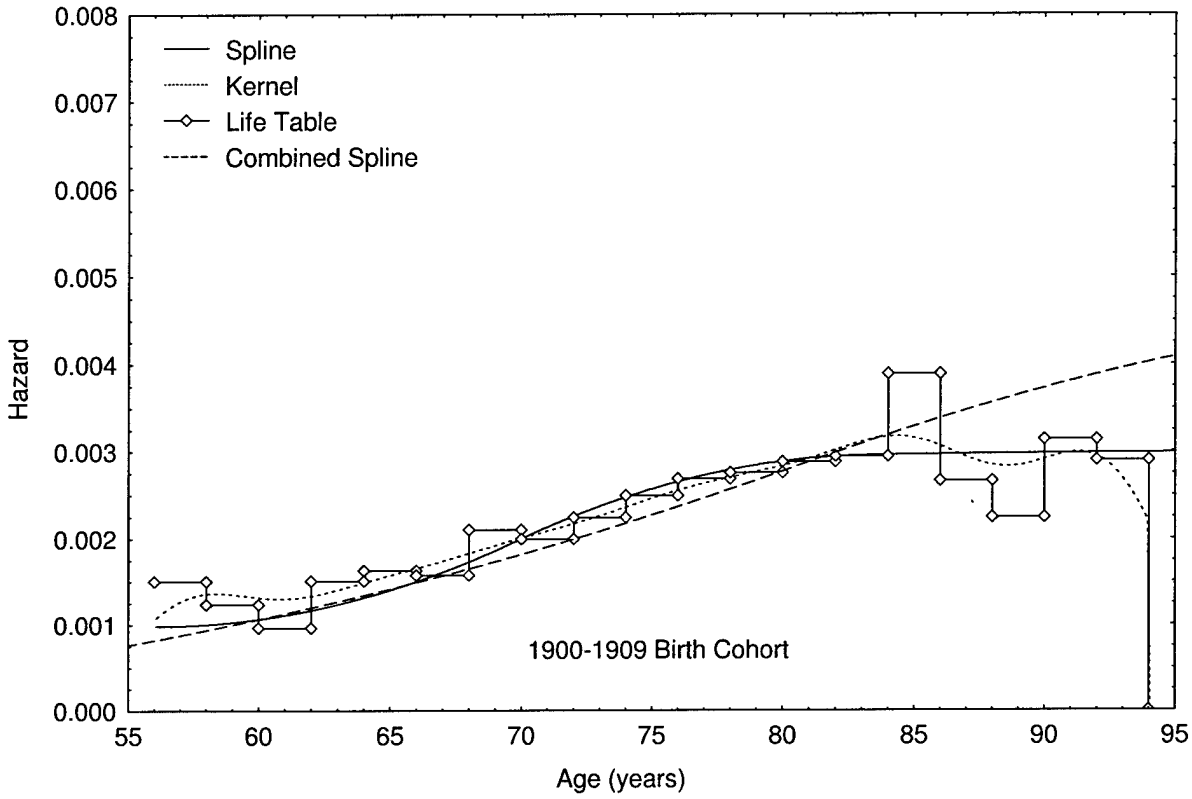

Spline
Kernel
Life Table
Combined Spline
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Hazard
1900-1909 Birth Cohort
55
60
65
70
75
80
85
90
95
Age (years)

FIGURE 1C

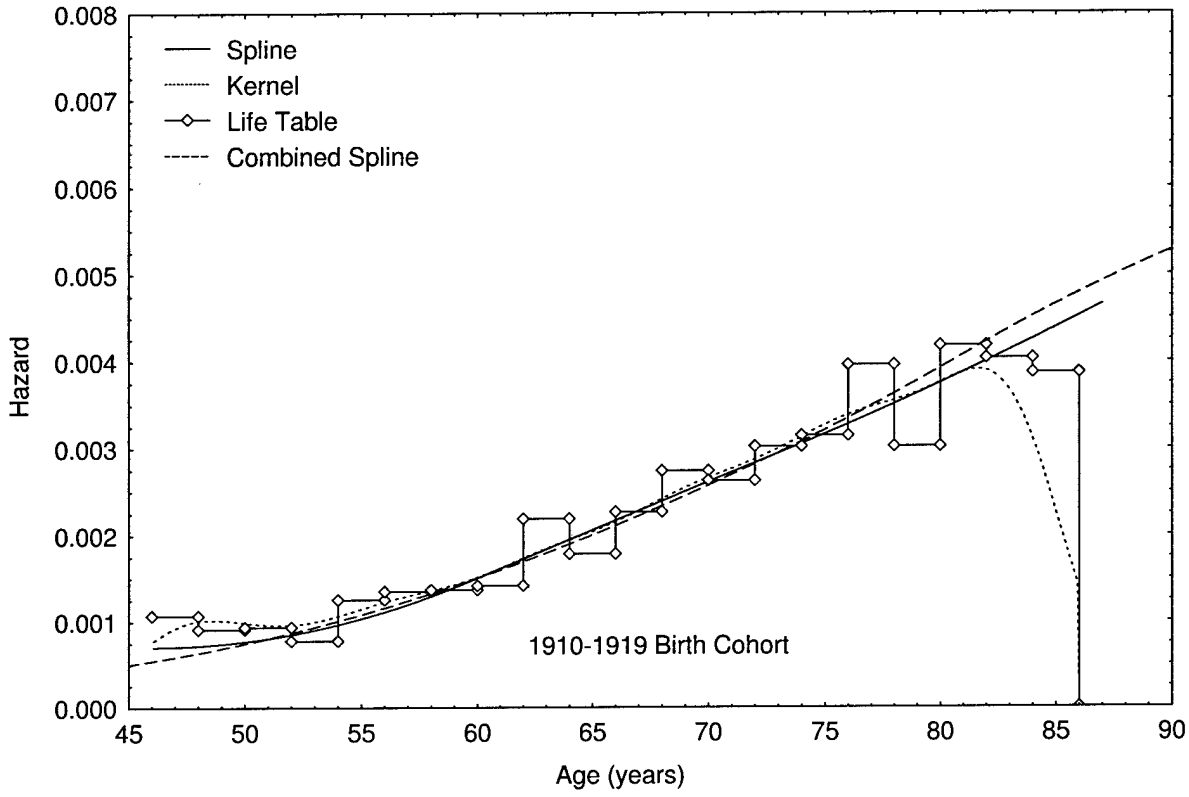

Spline
Kernel
Life Table
Combined Spline
Hazard
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
1910-1919 Birth Cohort
45
50
55
60
65
70
75
80
85
90
Age (years)

FIGURE 1D

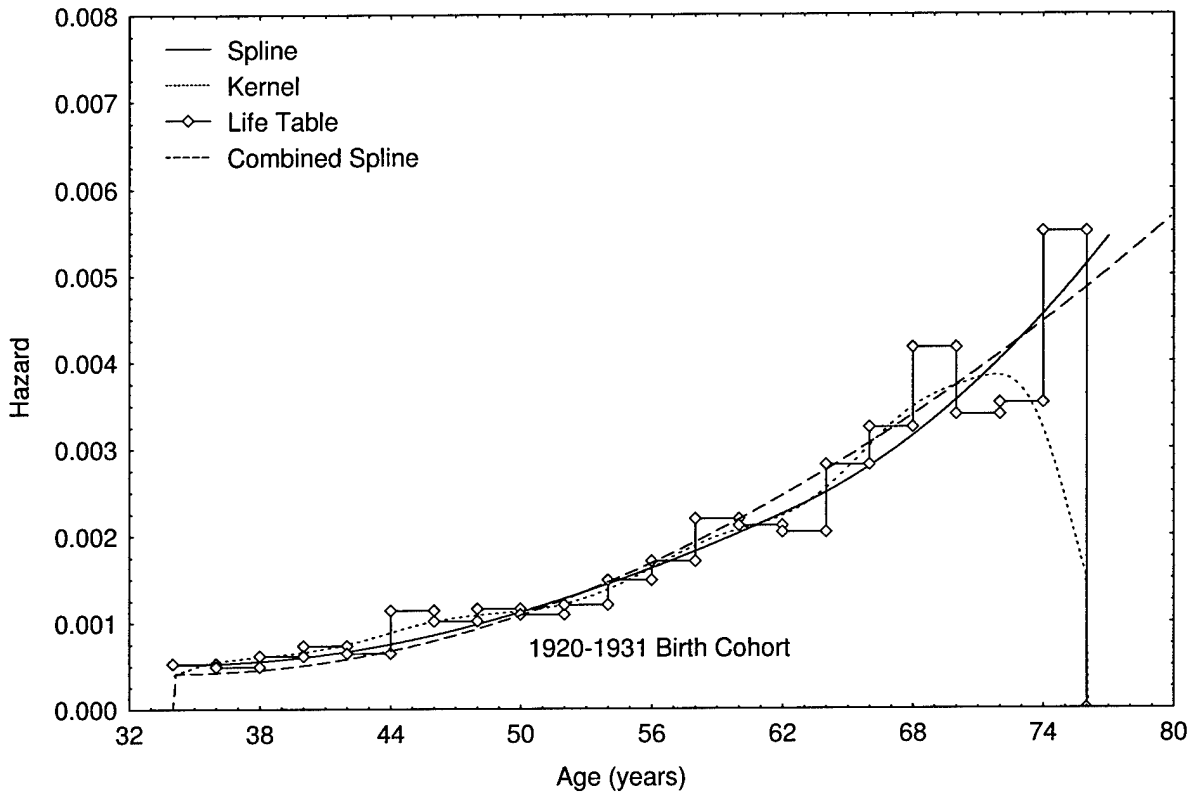

Spline
Kernel
Life Table
Combined Spline
Hazard
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
1920-1931 Birth Cohort
32
38
44
50
56
62
68
74
80
Age (years)

FIGURE 1E

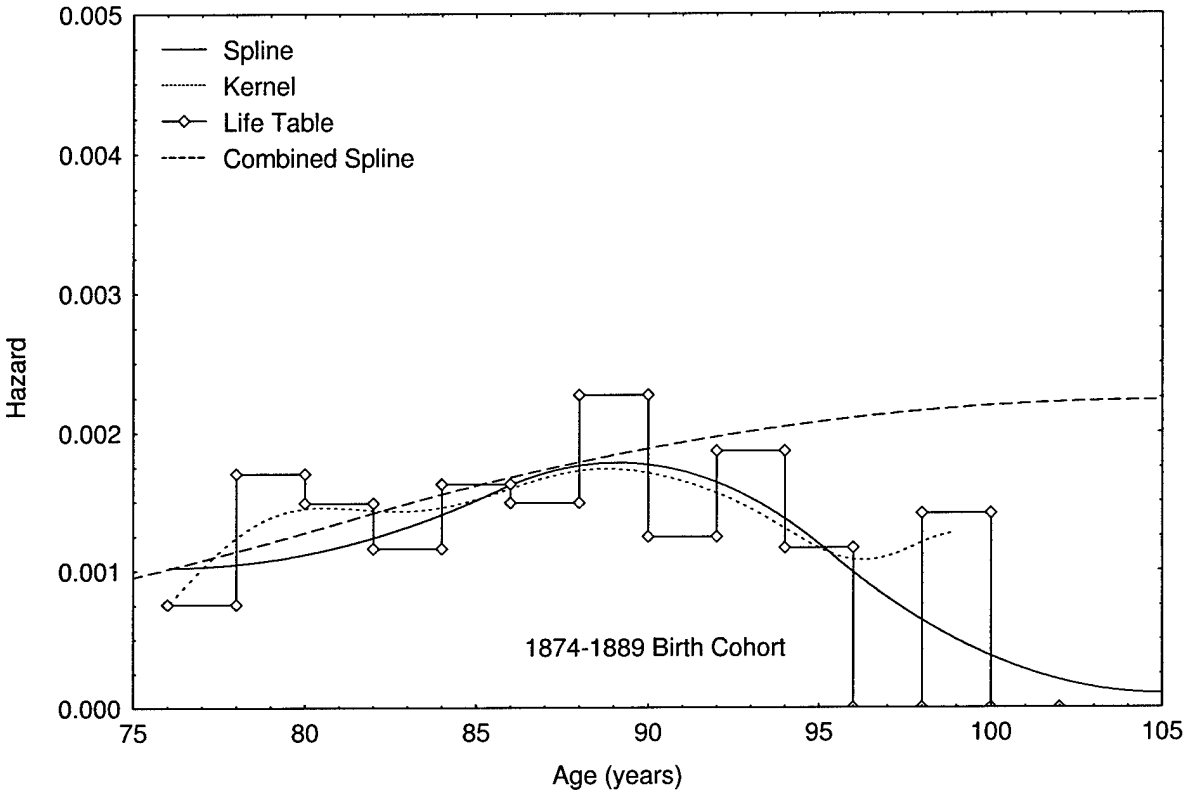

Spline
Kernel
Life Table
Combined Spline
0.005
0.004
0.003
0.002
0.001
0.000
Hazard
1874-1889 Birth Cohort
75
80
85
90
95
100
105
Age (years)

FIGURE 2A

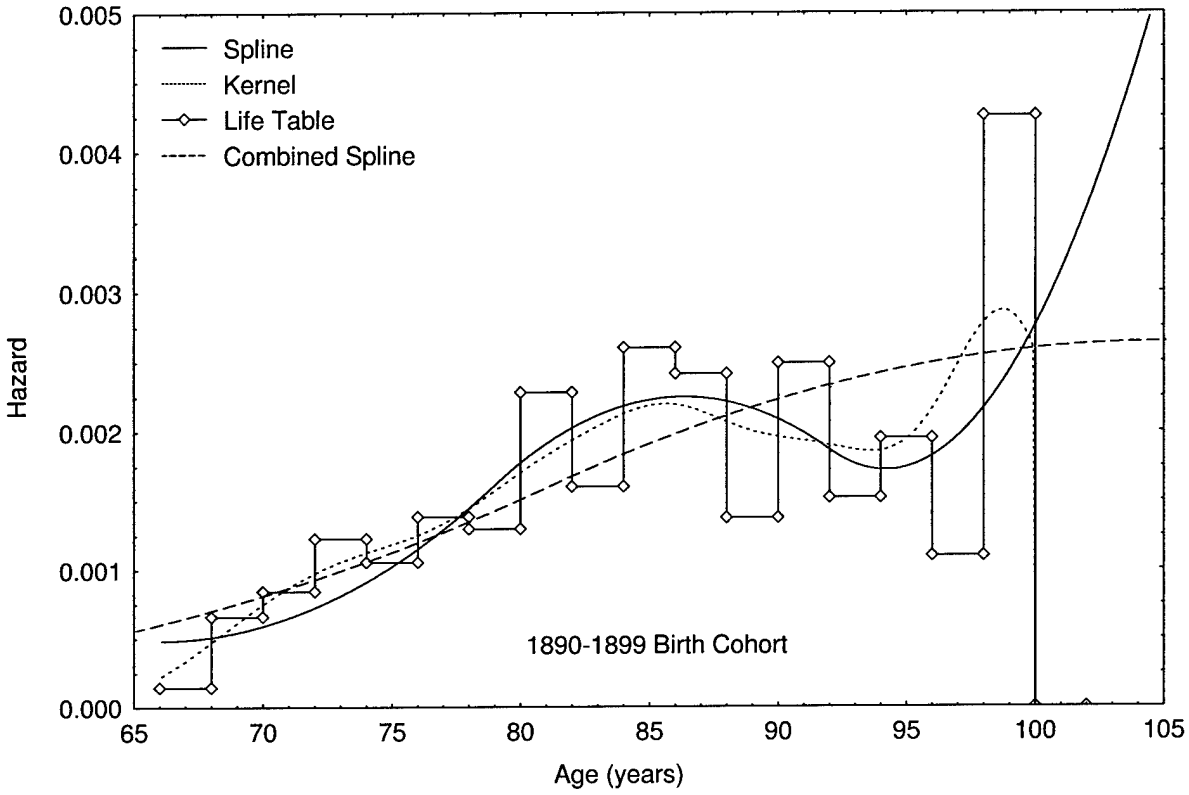

0.005
0.004
0.003
0.002
0.001
0.000
Hazard
Spline
Kernel
Life Table
Combined Spline
1890-1899 Birth Cohort
65
70
75
80
85
90
95
100
105
Age (years)

FIGURE 2B

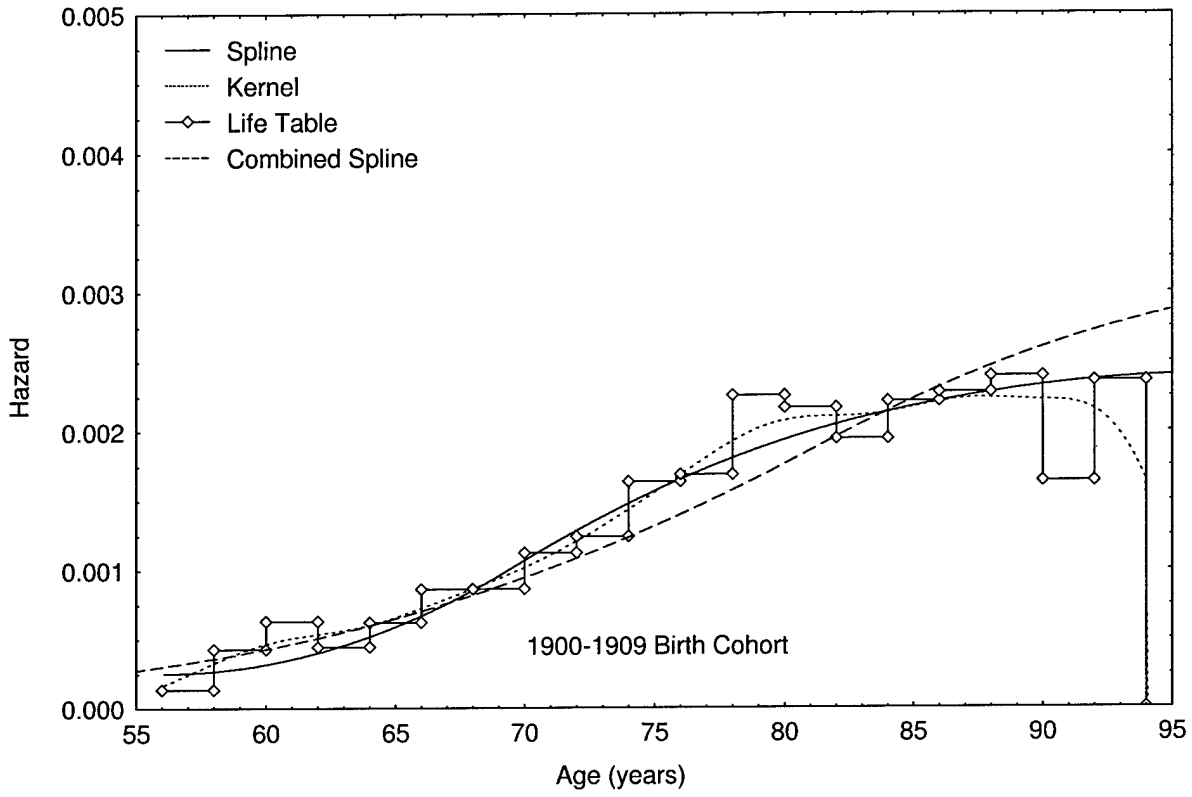

0.005
0.004
0.003
0.002
0.001
0.000
Hazard
Spline
Kernel
Life Table
Combined Spline
1900-1909 Birth Cohort
55
60
65
70
75
80
85
90
95
Age (years)

FIGURE 2C

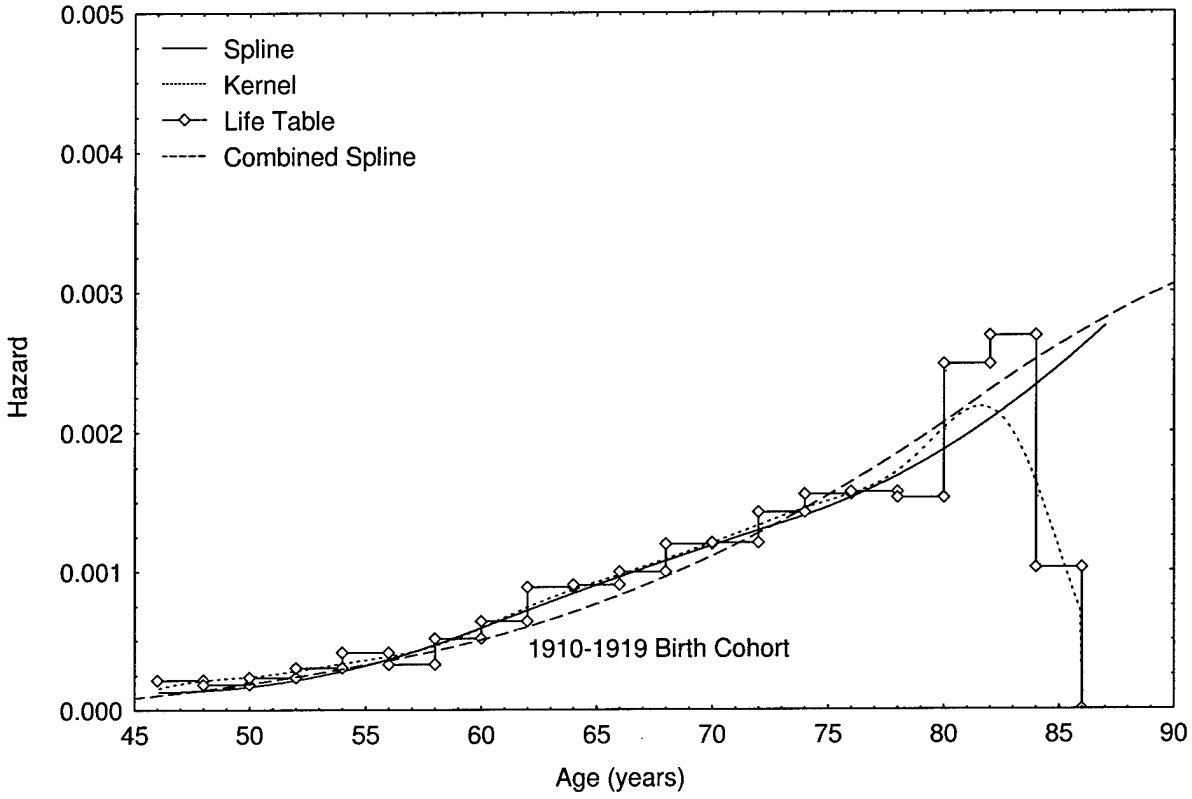

0.005
0.004
0.003
0.002
0.001
0.000
Hazard
Spline
Kernel
Life Table
Combined Spline
1910-1919 Birth Cohort
45
50
55
60
65
70
75
80
85
90
Age (years)

FIGURE 2D

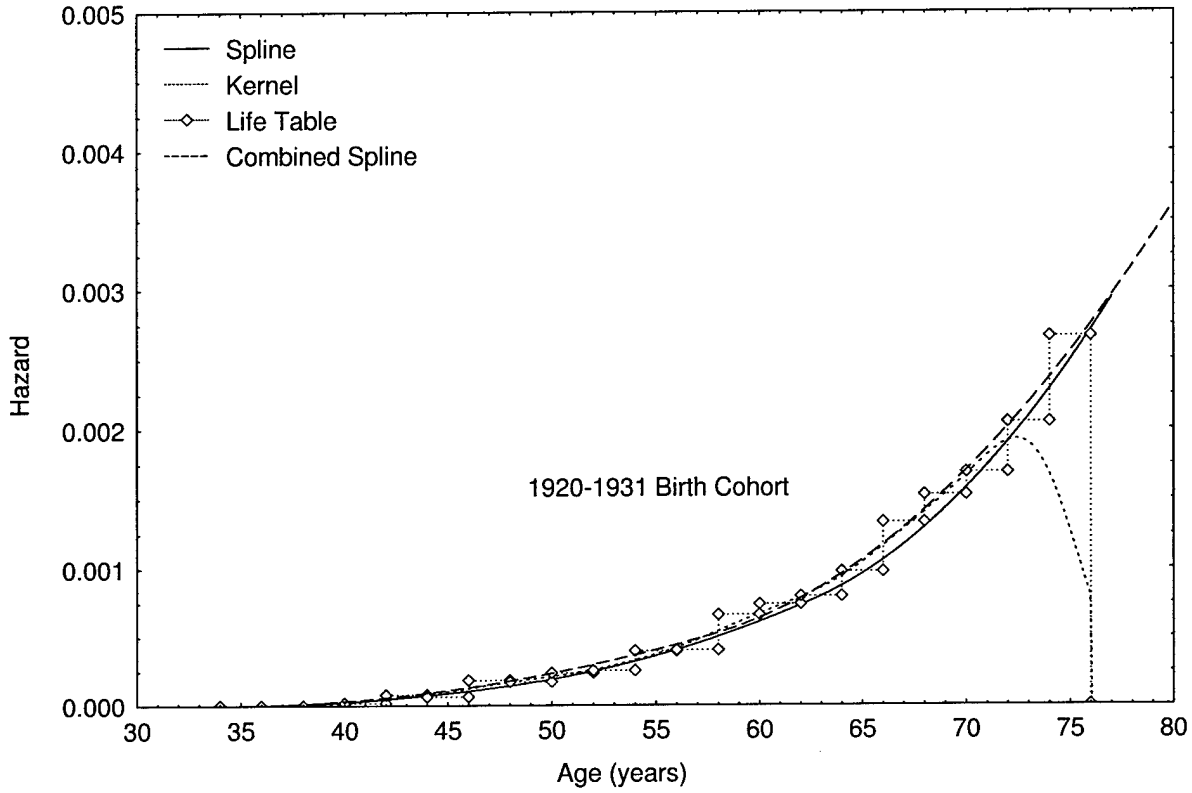

0.005
0.004
0.003
0.002
0.001
0.000
Hazard
Spline
Kernel
Life Table
Combined Spline
1920-1931 Birth Cohort
30
35
40
45
50
55
60
65
70
75
80
Age (years)

FIGURE 2E

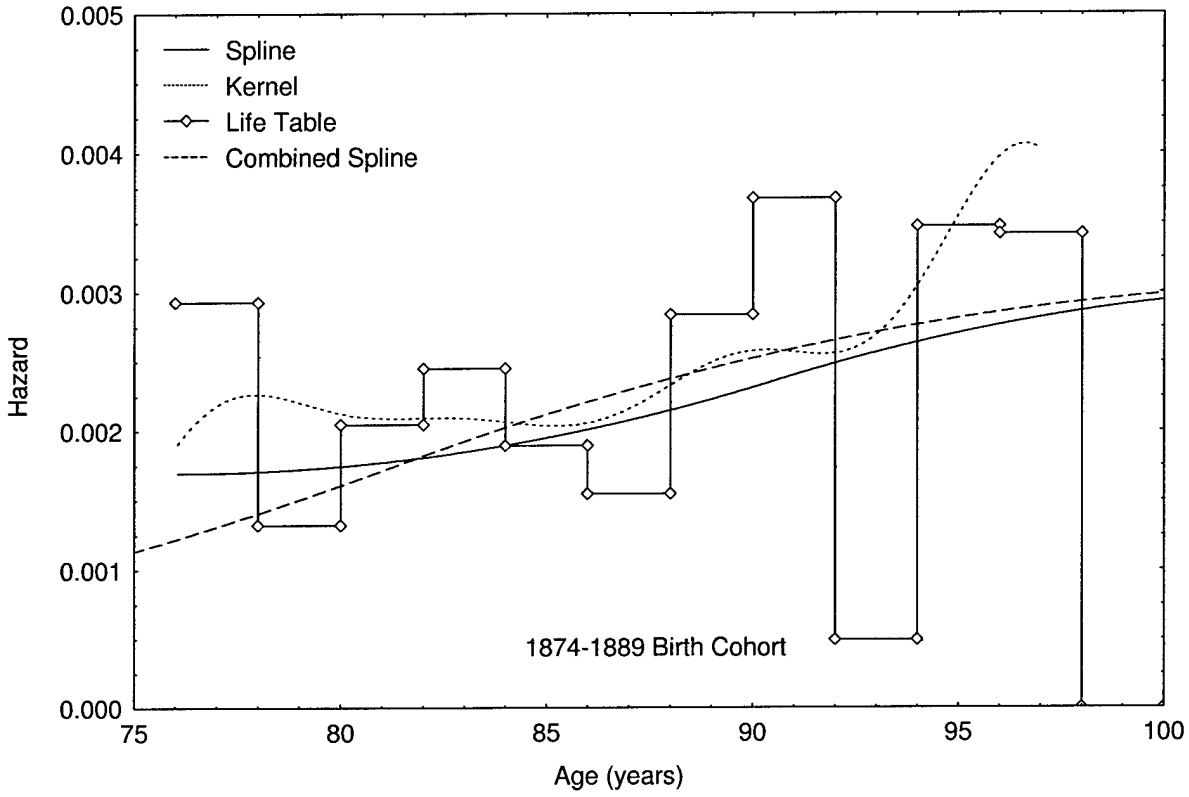

0.005
0.004
0.003
0.002
0.001
0.000
Hazard
Spline
Kernel
Life Table
Combined Spline
1874-1889 Birth Cohort
75
80
85
90
95
100
Age (years)

FIGURE 3A

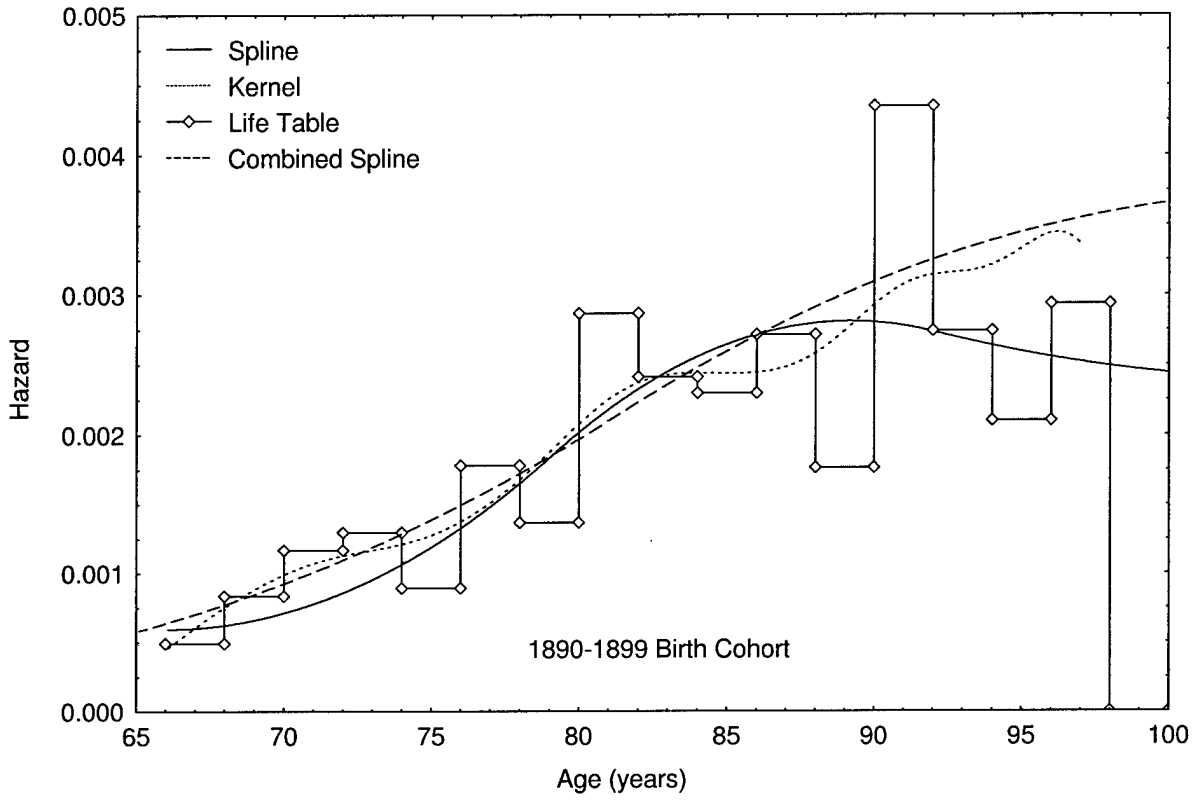

Spline
Kernel
Life Table
Combined Spline
0.005
0.004
0.003
0.002
0.001
0.000
Hazard
1890-1899 Birth Cohort
65
70
75
80
85
90
95
100
Age (years)

FIGURE 3B

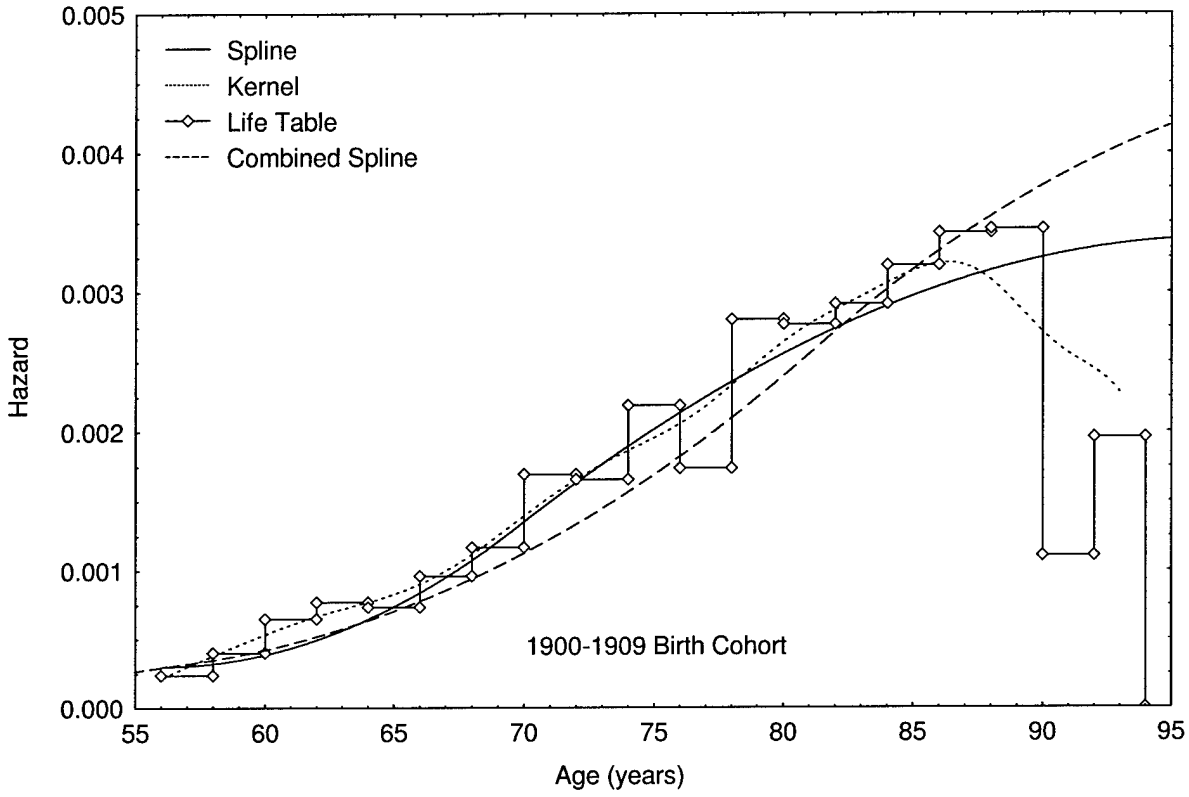

Spline
Kernel
Life Table
Combined Spline
0.005
0.004
0.003
0.002
0.001
0.000
Hazard
1900-1909 Birth Cohort
55
60
65
70
75
80
85
90
95
Age (years)

FIGURE 3C

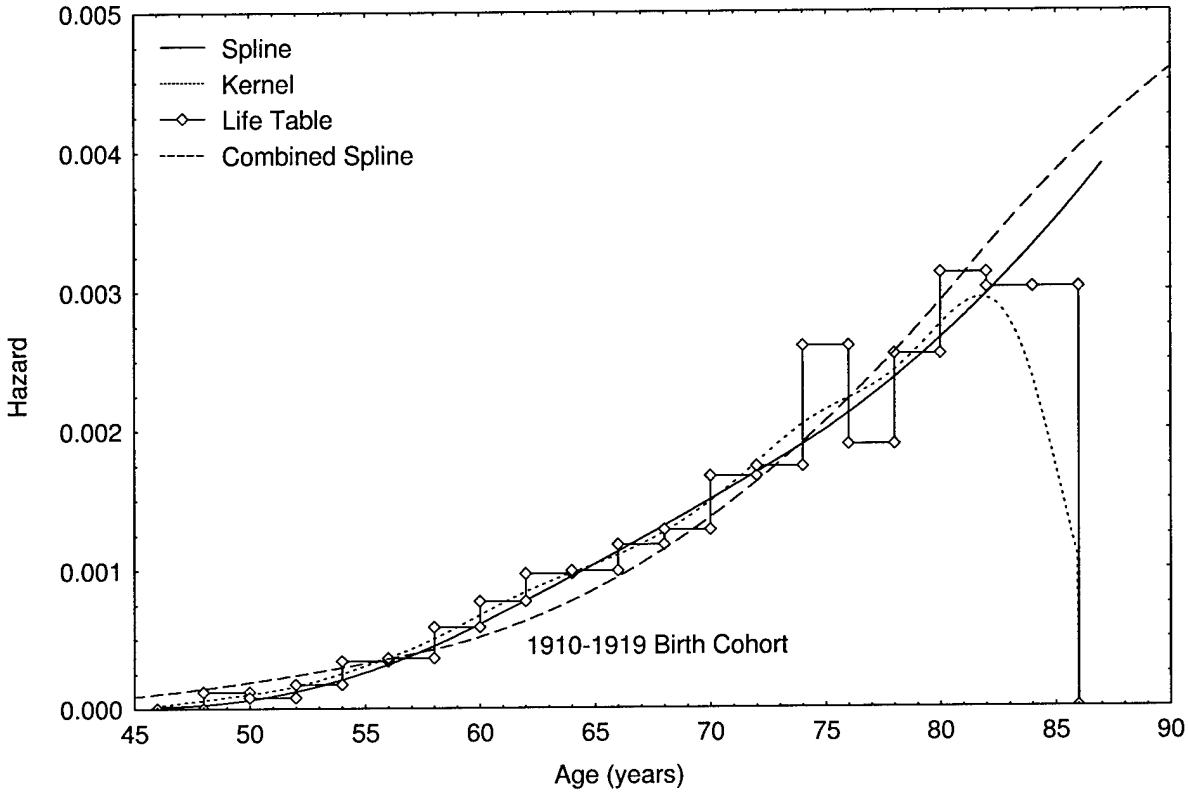

Spline
Kernel
Life Table
Combined Spline
0.005
0.004
0.003
0.002
0.001
0.000
Hazard
1910-1919 Birth Cohort
45
50
55
60
65
70
75
80
85
90
Age (years)

FIGURE 3D

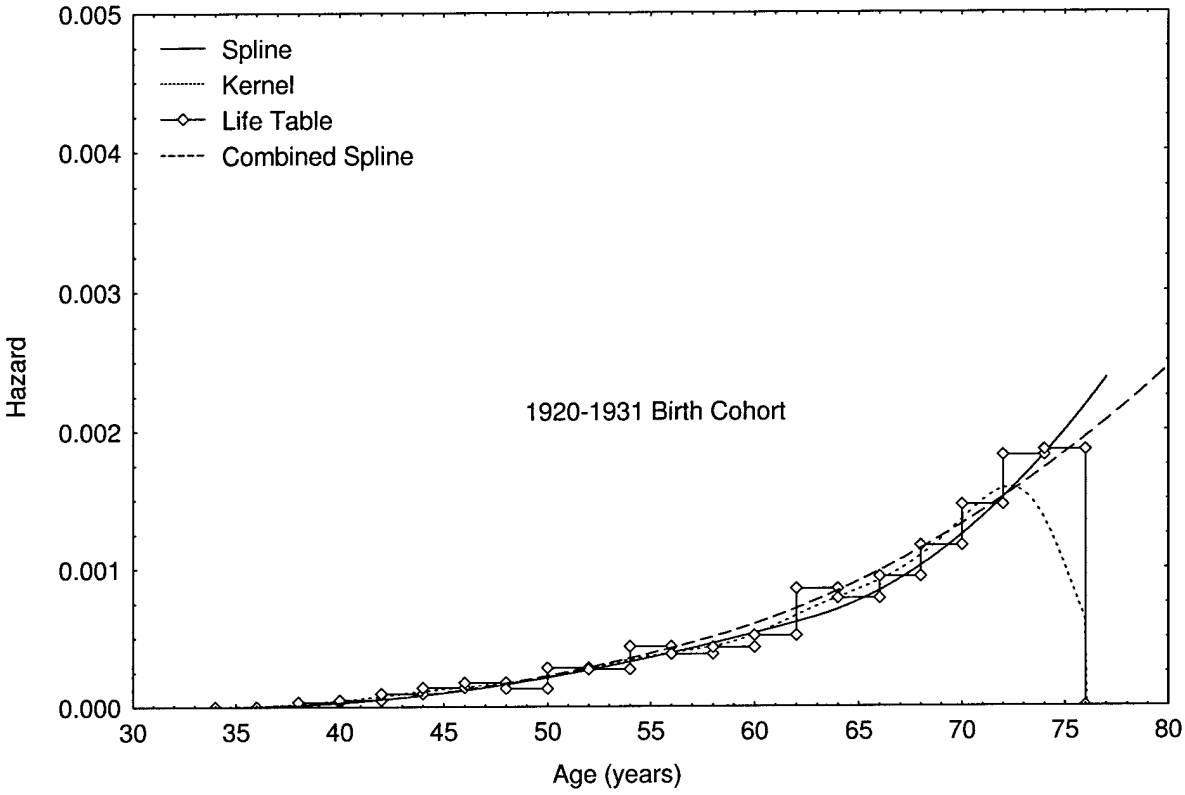

0.005
0.004
0.003
0.002
0.001
0.000
Hazard
Spline
Kernel
Life Table
Combined Spline
1920-1931 Birth Cohort
30
35
40
45
50
55
60
65
70
75
80
Age (years)

FIGURE 3E

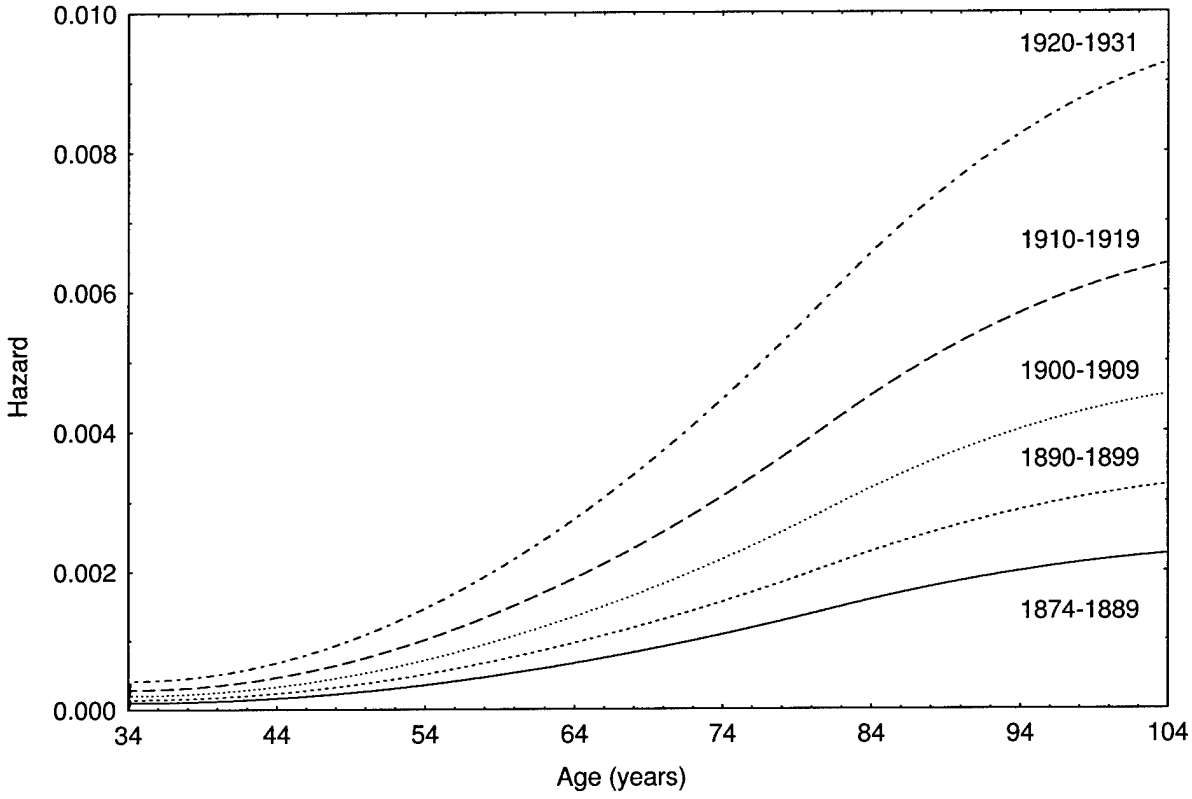

0.010
0.008
0.006
0.004
0.002
0.000
Hazard
34
44
54
64
74
84
94
104
Age (years)
1920-1931
1910-1919
1900-1909
1890-1899
1874-1889

FIGURE 4

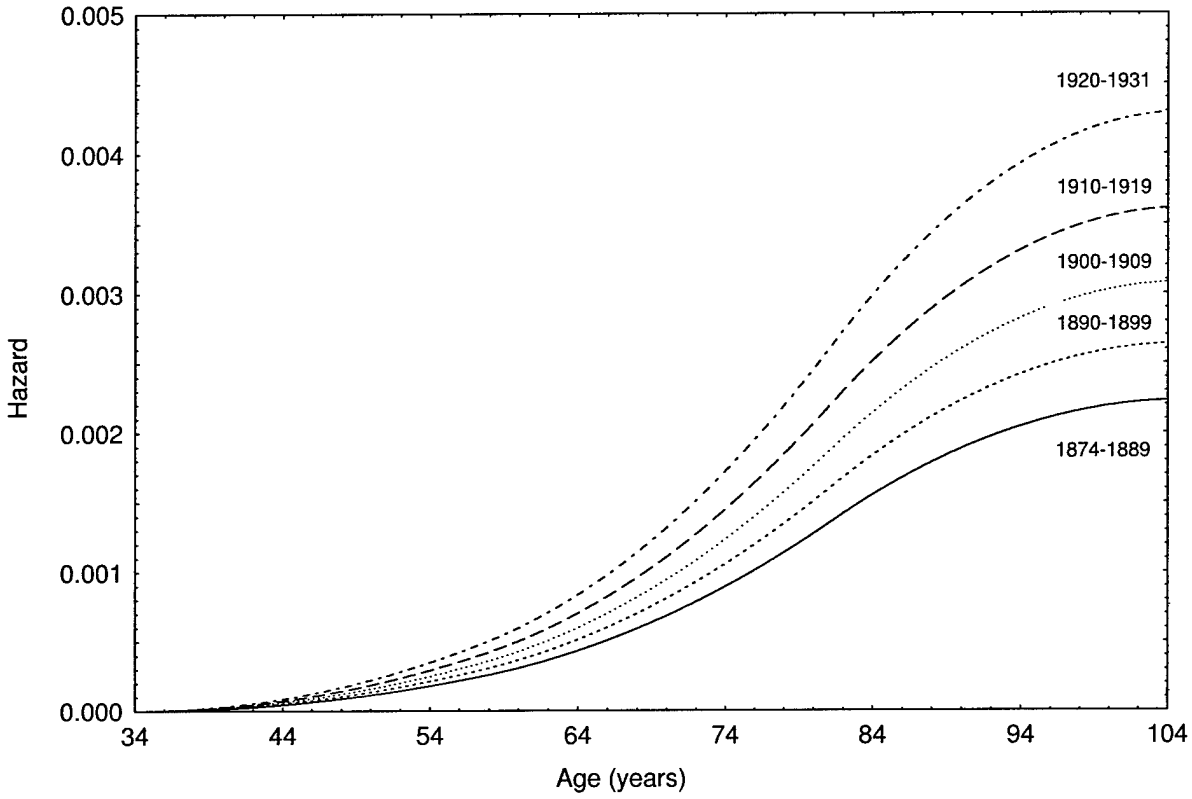

0.005
0.004
0.003
0.002
0.001
0.000
Hazard
34
44
54
64
74
84
94
104
Age (years)
1920-1931
1910-1919
1900-1909
1890-1899
1874-1889

FIGURE 5

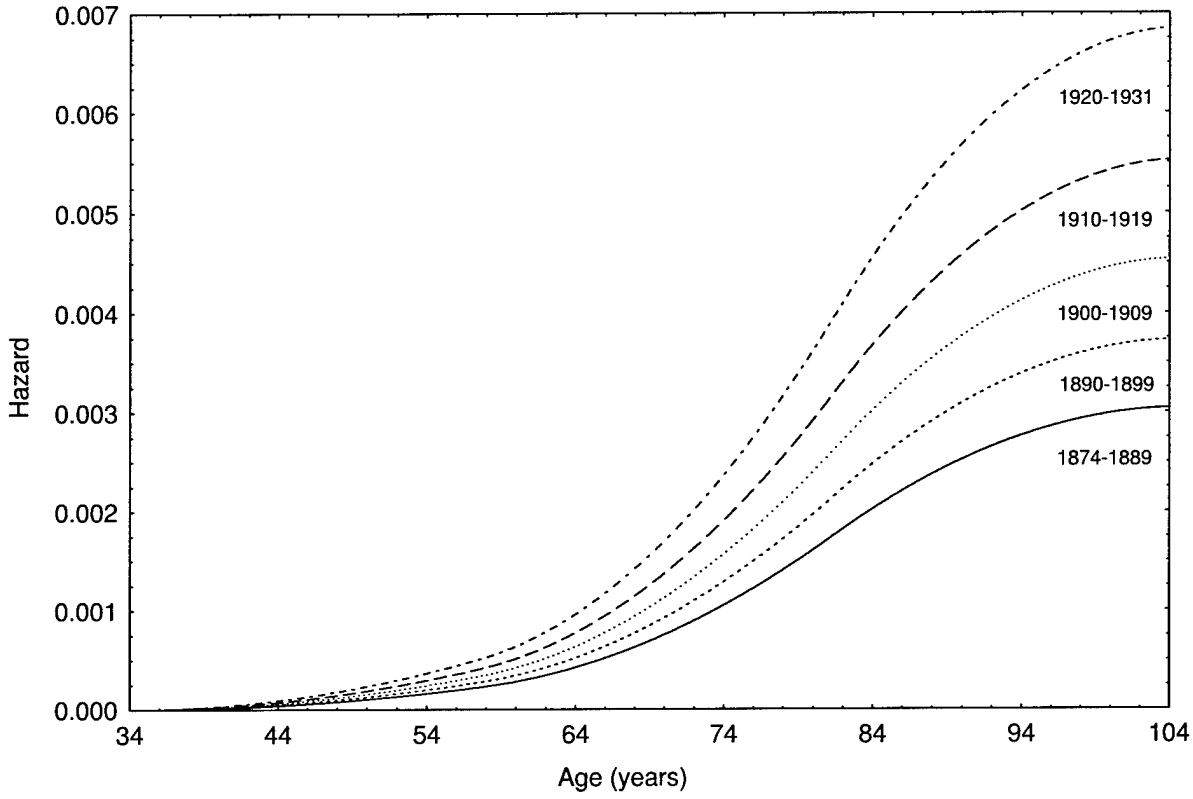

0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Hazard
34
44
54
64
74
84
94
104
Age (years)
1920-1931
1910-1919
1900-1909
1890-1899
1874-1889

FIGURE 6

Appendix 2

Measures of Familial Aggregation as Predictors of Breast Cancer Risk

Kenneth M. Boucher and Richard A. Kerber

Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, 2000 Circle of Hope, Salt Lake City, Utah 84112

Running title: **Familial Predictors of Breast Cancer Risk**

Corresponding author:
Kenneth M. Boucher, Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112, U.S.A.
Phone: 801-585-9544, FAX: 801-585-5357
e-mail: ken.boucher@hci.utah.edu
8/2000

ABSTRACT

BACKGROUND: Several measures of familial disease aggregation have been proposed, but only a few of these are designed to be implemented at the individual level. We evaluate four of them in the context of breast cancer incidence.

METHOD: A population-based cohort consisting of 114,429 women born between 1874 and 1931 and at risk for breast cancer after 1965 was identified by linking the Utah Population Data Base and the Utah Cancer Registry. Three competing methods were used to obtain predictors of familial aggregation of risk: the number of first degree relatives with breast cancer, the posterior probability of carrying BRCA1 or BRCA2, and the Familial Standardized Incidence Ratio (FSIR), which weights the disease status of relatives based on their degree of relatedness with the proband. Spline regression methods were used to estimate the hazard function, stratified by measures of familial aggregation.

RESULTS: When the measures of family history are dichotomized with approximately 8.5% of subjects in the high risk category, presence of a first degree relative and FSIR perform equally well at determining individual risk, with the high risk category having approximately twice the risk at all ages. The posterior probability of BRCA1 and BRCA2 performed less well. When FSIR is further stratified, the top 0.1% have an approximate 4-fold increase in risk. The risk appears to be increasing through all age groups.

CONCLUSIONS: Family history is a highly significant indicator of risk for breast cancer.

KEYWORDS: familial risk, hazard function, truncation, survival analysis, breast cancer

Introduction

Heterogeneity in a population may lead to population estimates of the hazard that do not reflect individual risk. For example, if we let $\lambda(t)$ denote the hazard function, and p the probability of immunity to a particular disease, it follows from the formula

$$p = \lim_{t\to\infty} \exp\left\{-\int_0^t \lambda(u)du\right\},$$

that there are individuals who are "immune" in the population exactly when the hazard function has finite integral. In particular, $\lim_{t\to\infty} \lambda(t) = 0$, provided the limit exists. More generally, a large degree of heterogeneity in disease susceptibility may lead to a population hazard function with one or more well-defined maxima. The maxima may correspond to discrete subpopulations with different genetic predisposition to disease. A maximum may also result from a continuous frailty, as the surviving population at higher ages may be overrepresented by individuals with lower risk[1].

In fact, there is evidence of heterogeneity for most cancers. According to Easton[2], "All cancer types exhibit familial clustering, suggestive of a significant inherited component". He goes on to conclude that as of 1994 known cancer genes accounted for 0.5-1% of all cancer cases, and that this figure would increase as more cancer genes are discovered. The breast cancer genes BRCA1 and BRCA2 both contribute to an increased risk of breast cancer. BRCA1 has an estimated allele frequency of between 0.0002 and 0.001 (95% CI)[3], and accounts for about 3% of diagnosed breast cancer[4]. The allele frequency of mutations in BRCA2 is estimated at 0.00022[5]. Vehmanen *et al.*[6] found that only 21% of breast cancer families were accounted for by mutations of BRCA1 and BRCA2, providing indirect evidence for the existence of other, undiscovered breast cancer genes.

In our previous paper[7], linked populations-based data from the Utah Cancer Registry and the Utah Population Data Base was used to estimate the population-level hazard function for breast and colorectal cancer, stratified by birth cohort. We found that the hazard functions for both breast and colorectal cancer appeared to be monotone increasing functions for both genders and all birth cohorts. This contrasts with the model-based estimates of Moolgavkar *et al.*[8], who found the hazard function to sharply decrease starting sometime past the age of 70.

The lack of clear multiple modes in the hazard function made it clear that more delicate methods would be needed to account for the known heterogeneity of risk.

A number of measures of familial disease aggregation have been used or proposed, but only a few of these are designed to be implemented at the individual level. The most common epidemiologic measure of familial risk is an indicator of whether one or more first-degree relatives has been diagnosed with cancer or some other disease. Khoury and Flanders[9] have noted that measures of this sort are prone to bias under a variety of conditions. Nonetheless, it is a widely used and easily understood measure of familial risk that can easily be ascertained in a clinical setting. A second category of family history measures suggested by Kerber[10] are derived from the complete risk experience of all observable biological relatives adjusted for the age, sex, number and degree of the relatives. The total familial risk is summarized as a familial standardized incidence ratio (FSIR) or a familial rate (FR). FSIR and FR are less prone to bias and substantially more sensitive than a crude indicator variable, but require fairly detailed family history data which may rarely be available in a clinical setting. A third measure, particularly relevant for breast cancer, was introduced by Parmigiani *et al.*[12]. Parmigiani et al. estimated the posterior probability that an individual carried the breast cancer genes BRCA1 and BRCA2 using information on first and second degree relatives of the subject. The method relies heavily on prior estimates of risk to carriers, and prior estimates of prevalence of the the genes.

In this paper age specific estimates of the hazard function for breast cancer incidence is estimated, stratified by the above measures of family history. It is found that FSIR and presence of a first degree relative with breast cancer are highly significant predictors of increased risk, with an identified high risk category having twice the risk. The hazard function for breast cancer appears to increasing as a function of age in all risk groups.

Hazard Function Estimation

Data

The data for this study were obtained by linking records from the Utah Population Data Base (UPDB) with the Utah Cancer Registry (UCR). The UPDB consists of

the genealogical records of more than 1,000,000 individuals who were born, died, or married in Utah, or en route to Utah during the nineteenth and twentieth centuries. The available follow-up information comes either from Utah death certificates, which have been linked to the UPDB genealogical data every year from 1933 through the beginning of 1997, or from linkage of the HCFA beneficiary data to the UPDB. The study population consists of 122,208 women recorded in the Utah Population Database, who were born from 1874 to 1931 and for whom follow-up information is available that places them in Utah during the years of operation of the Utah Cancer Registry (1966-present). Subjects with purported follow-up past age 105 were excluded from the data. Potential subjects who had no relatives who were also in the risk set, and therefore for whom no measures of familial aggregation could be computed, were removed from the data. Excluding these two groups removed an additional 7779 women, leaving a study population of 114,429 women. There are 5,092 cases of female breast cancer in the data. Only female breast cancer was analyzed. Additional details on the data are given in Boucher and Kerber[7].

Nonparametric Hazard Estimation

The data described above are subject to random truncation: cases which occurred during or before 1965 are not recorded in the dataset. Subject were between the ages of 34 and 86, at the time of truncation. Thus, analysis of the data must take into account not only to the effects of right censoring, but also the effects of left truncation due to delayed entry into the risk set.

Let the truncation time Y have distribution function $G(y)$, the minimum of the failure and censoring time be X and have distribution function $F(x)$, and δ be the censoring indicator, with $\delta = 1$ signifying a censored observation. We require that truncation and censoring be independent of failure. Observations are conditional on $X > Y$. Let $G^*(y)$ and $F^*(x)$ be the corresponding distribution functions, conditional on $X > Y$. Let $S(x)$ be the survivor function for the failure time distribution. Suppose that we have observations $(Y_1^*, X_1^*, \delta_1^*), \ldots, (Y_n^*, X_n^*, \delta_n^*)$, from the conditional distribution, where for simplicity we describe the situation with no tied failure times. Our nonparametric methods are based on the nonparametric maximum like-

lihood estimator (NPMLE)

$$\hat{S}(t) = \prod_{X_i^* \leq t, \delta_i = 0} \left(1 - \frac{1}{R(X_i^*)}\right), \tag{1}$$

where $R(U) = \#\{Y_i^* < U \leq X_i^*\}$ is the number of subjects at risk at U, as described, for example in Keiding[12]. The Nelson-Aalen estimator of the cumulative hazard is given by

$$\hat{\Lambda}(t) = \sum_{X_i^* \leq t} R(X_i^*)^{-1}. \tag{2}$$

Parametric Hazard Estimation

We again assume that X, Y and δ are as above. We wish to have a parameterization $F_1(x, \vec{s}; \vec{z})$ of the failure time distribution, with covariate vector $\vec{s}$ and parameter vector $\vec{z}$. We denote the corresponding density $f_1(x, \vec{s}; \vec{z})$ and survival function $S_1(x, \vec{s}; \vec{z})$. We condition on $X > Y$, and in analogous fashion to what is done in the nonparametric setting, maximize the logarithm of the conditional likelihood. Let $\lambda(x, \vec{s}; \vec{z})$ denotes the hazard associated with $F_1(x, \vec{s}; \vec{z})$ and $\Lambda(x, \vec{s}; \vec{z})$ the cumulative hazard. The likelihood, conditional of $X > Y$, becomes

$$\log(CL) = \sum_{i=1}^{n} [\delta_i \log(\lambda(x_i, \vec{s}_i; \vec{z}_i)) - (\Lambda(x_i, \vec{s}_i; \vec{z}) - \Lambda(y_i, \vec{s}_i; \vec{z}_i))]. \tag{3}$$

We modeled the hazard via quadratic splines[13]. A quadratic spline with m knots specifies the hazard to be of the form

$$h_m(t) = \sum_{i=0}^{2} \gamma_{0i} t^k + \sum_{j=1}^{m} \gamma_{j2} (t - \tau_j)_+^2 \tag{4}$$

where $(x)_+ = \max(x, 0)$. For each birth cohort, we fit splines with knots which were equally spaced in the interior of the interior $[T_{min}, T_{max}]$, where T_{min} is the minimum truncation age in the cohort and T_{max} the maximum follow-up (failure or censoring) time. Restrictions were placed on the coefficients to ensure that $\lambda_m(t)$ remained positive for all t. Thus with m knots the number of parameters was $m + 3$. Models were fit by maximizing the conditional likelihood.

We fit proportional hazards models with splines $\lambda_m(t)$ for the baseline hazard and a covariate vector $\vec{s}$ with one component for birth year and perhaps one component for family history. Birth year was shown to be highly significant in our previous

paper[7], and may account for such effects as a decrease in parity and an increase in the efficacy of detection methods with time. The resulting hazard function has the form

$$\lambda_m(t, \vec{s}; \vec{\beta}) = \exp(\sum_{j=1}^{2} \beta_j s_j) h_m(t). \tag{5}$$

The model was fit using the conditional likelihood (3) with $\lambda(x, \vec{s}, \vec{z}) = \lambda_m(x, \vec{s}; \vec{\beta})$.

The hazard function was estimated for female breast cancer. The spline estimates were computed by maximizing $\log(CL)$ using the algorithm of Powell[14]. We started with one knot and increased the number of knots until the fit was not improved, as determined by the likelihood ratio test at the significance level $\alpha = 0.05$. The life table estimator based on (2) was used for comparison with the spline-based estimator.

Methods of Familial Aggregation

Number of First Degree Relatives

The simplest and most easily understandable is the number of first degree relatives with breast cancer. Of the 114,429 women in the data set, 9765, or approximately 8.5%, had at least one first degree relative with breast cancer also represented in the data, and 795 women, or 0.69%, had two or more relatives in the data. Having more than two first degree relatives with breast cancer was extremely rare: 56 women had three, and 10 women had the maximum of four.

Posterior Probability of BRCA1 and BRCA2

We used the method of Parmigiani et al[11], and implemented in the computer program BRCAPRO, available from the authors, to computed posterior probabilities of carrying BRCA1 and BRCA2 mutations for each of our subjects. The method uses age at onset of breast and ovarian cancer for first and second degree relatives to compute posterior probabilities of carrying BRCA1 and BRCA2. The method incorporates prior distributions for the risk of breast and ovarian cancer to carriers and noncarriers of the breast cancer genes BRCA1 and BRCA2 as well as prior estimates of distribution of the population level carrier probabilities. We used the prior probability distributions suggested by Parmigiani *et al.*[11].

Were were able to compute posterior probabilities for 114,221 (or 99.8%) of the subjects with a first degree relative in the database. The mean carrier probabilities were 0.000301 and 000098 for BRCA1 and BRCA2 respectively, with medians of 0.000098 and 000015. The distributions of the posterior carrier probabilities are shown in Figures 1 and 2.

Familial Standardized Incidence Ratio

The second measure of familial aggregation is a modification of the familial standardized incidence method (FSIR)[10]. The familial standardized incidence ratio is derived from the complete risk experience of all observable biological relatives, adjusted for age, sex, number and degree of the relatives. FSIR is defined in terms of the kinship coefficient[15] $c(i,j)$ between individuals i and j, which gives the probability that two individuals share a gene at a given locus. The kinship coefficient is defined by $c(i,j) = (1/2)\sum_{p=1}^{P_{i,j}} 2^{-l(p)}$, where $P_{i,j}$ is the total number of paths between individuals i and j, and $l(p)$ is the length in reproductive events of each path p. Let $I_j = 1$ if the jth member has the disease and 0 otherwise. Finally, we suppose that we have a stratified population, the population incidence in the kth statum is given by λ_k, and let t_{jk} be the time that the jth person spent in the kth stratum of risk. The familial standardized incidence ratio is then defined, for the i individual, by

$$FSIR_i = \frac{\sum_{j=1}^{J} I_j c(i,j)}{\sum_{k=1}^{K}\sum_{j=1}^{J} t_{jk}\lambda_k c(i,j)}$$

In deriving a measure of variance VAR_i for $FSIR_i$, it was assumed that the denominator of the above expression is fixed, and that for each fixed path length the number of observed cases follows a Poisson distribution with mean equal to the expected number of cases in the stratum. The population risk estimates used to construct the denominator of $FSIR_i$ were assumed to be fixed.

A difficulty with using the "raw" FSIR scores is that the amount of information from which it is constructed for a particular individual is highly variable. A low FSIR score could be an indicator of low risk or simply reflect small family size. We therefore chose to adjust the scores using an empirical Bayes procedure before incorporating them into a regression analysis. As the raw FSIR scores are highly skewed, we first transformed them using a loglog transform $\text{loglog}(FSIR) = \log(1+\log(1+FSIR))$.

The basic assumption of the empirical Bayes adjustment is that the "true" values μ of loglog($FSIR$) are normally distributed. The mean and variance of μ are estimated empirically and iteratively from the data. The procedure we use is similar to the one suggested by Greenland and Robins[16].

More specifically, we suppose that after iteration $n-1$ we have current estimates $\mu_{i,n-1}$ and $\sigma^2_{i,n-1}$ for the true value and ith individual, as well as an overall mean μ_{n-1} and variance σ^2_{n-1} for the μ_i. We then computed new estimates using the formulas

$$\mu_{i,n} = \mu_{n-1} + (\frac{\sigma^2_{n-1}}{\sigma^2_{n-1} + \sigma^2_{i,n-1}})(Y_i - \mu_{n-1}),$$

where $Y_i = \text{loglog}FSIR_i$, and with variance estimated by

$$\sigma^2_{i,n} = \frac{VAR_i}{(\exp(\mu_{i,n-1})\exp(\exp(\mu_{i,n-1}) - 1))^2}$$

given by the delta method. We then computed the sample mean and variance of $\mu_{i,n}$, over all the subjects to get μ_n and σ^2_n.

The distribution of loglog(FSIR), before and after transformation, are displayed in Figure 1. Note that the "raw" distribution is bimodal, with a mode at zero which disappears after transformation.

Results

Dichotomized Comparison of Familial Risk

We dichotomized each of our measures of famililial risk, with the high risk category representing approximately 8.5% of the data in each case. This was a natural cut point, as it represents the proportion of subjects with one or more first degree relatives with breast cancer. The cutoff for FSIR roughly corresponds to a relative risk of two to family members. The cut points for the posterior probability of BRCA1 and BRCA2 come at points where the posterior probability is rather small, less than 0.0005 in both cases. The number of subjects in each category and the ranges for the variables are presented in Table 1.

Our previous analysis indicated that a highly significant birth-year effect exists in the data[7], with a women born ten years later having an estimated 40% increased age-specific risk. Birth-year was included as an additional covariate in all regression

analyses. The baseline risk was estimated using splines, with the proportional hazards model used for birth-year and familial risk. As with most of the models, we found that two knots were sufficient to provide an optimal fit. Separate estimates of the age-specific hazard for each level of each of our familial risk measures are presented in Figure 4. For comparison we provided life table estimates of the risk. The life table estimates are not adjusted for birth-year. The life table estimates are flatter, and this may be explained by a significant birth-cohort effect. Subjects contribute to the risk estimates only for a period of at most 33 years of their lives, namely the period from 1965-1998. A women born in 1890 contributes only after age 75, while a women born in 1930 contributes from age 35 until the age of 68.

The presence of a first degree relative with breast cancer and the dichotomized FSIR variable each appear to be equally effective at distinguishing high risk subjects, with the high risk category having about double the risk, while the posterior probability of BRCA1 and BRCA2 appear to be less effective.

We performed a more detailed stratified analysis of FSIR. The category boundaries were the approximate 75th, 90th, and 99.9th percentiles of the (adjusted) FSIR distribution. The upper category roughly corresponds to the reported fraction of the general population carrying known breast cancer genes. The number of subjects, cases, and category boundaries are given in Table 2. Bootstrap confidence bands were computed as well as an indicator of the reliability of the estimates. The estimates of the age-specific hazard and percentile-based bootstrap confidence intervals are presented in Figure 5. The bootstrap confidence intervals are based on 100 bootstrap samples, except for the ¡75th percentile category, which is based on 20 bootstrap samples, because of the extensive time it took to fit the models to the large datasets.

Regression Methods Incorporating Familial Risk as a Covariate

We incorporated the posterior probabilities of BRCA1 and BRCA2 and their logarithms,as well as loglog$FSIR$ as continuous variables in separate analyses, using a proportional hazards model with birth-year as an additional covariate. The log-likelihoods and the values of χ^2_1 are presented in Table 3. We see that the best result (in terms of statistical significance) is obtained by including the loglog$FSIR$, where

we get a likelihood ratio $\chi^2_1 = 316.72, (p < 0.00001)$.

We also considered the indicator variable NFIRST for presence/absence of a first degree relative, in a proportional hazards model. From Figure 4A it can be seen that the proportional hazards assumption is not grossly violated. The variable NFIRST was highly significant (likelihood ratio $\chi^2_1 = 185.6, p < 0.0001$). Addition of a second indicator variable for two or more first degree relatives with breast cancer did not improve the likelihood significantly (data not shown).

Discussion

We have applied several methods of measuring familial aggregation at the individual level to breast cancer data. All prove to be signficantly significant predictors of individual risk. Judging by the difference in risk estimates, as well as the likelihood ratio test, presence of a first degree relative and FSIR appear to be better indicators of increased risk than the posterior probability of BRCA1 or BRCA2. Judging solely by the likelihood ratio test, one would prefer FSIR.

FSIR may be thought as an extension of the cruder number of first degree relatives with breast cancer, adjusting for the level of relatedness and expected disease. It is therefore not surprising to find that it performs better.

Although the estimates become less reliable at increasing age, the hazard function for breast cancer appears to be essentially non-decreasing in all the categories of all familial measures considered. Thus we find no evidence of an "immune fraction" in this analysis. The curves for different levels of risk appear not to merge or cross, indicating that the increased risk to those with a family history does not dissipate after a certain age.

Other investigators have either estimated or simply assumed that the risk of breast cancer decreases past a certain age. As previously noted, Moolgavkar *et al.*[8], found the hazard function to sharply decrease starting sometime past the age of 70. By age 90, the risk has decreased to about 1/3 of the peak. Parmigiani *et al.*[11] fit breast cancer incidence data from Easton *et al.*[17] to a three parameter gamma distribution. Implicit in this fitting procedure is the assumption that the risk to carriers of BRCA1 and BRCA2 decreases to zero with age. There is little actual

evidence for this in the fitted data, as the last age is 70. Although based on sparse data, our estimates show no evidence for decreased risk to carriers at advanced age. It may be important for further modeling efforts to better understand the hazards to carriers of disease susceptability genes, particularly at more advanced ages, where data are sparse.

Acknowledgments

This research was supported, in part, by NCI Cancer Center Support Grant 2P30 CA 42014, U.S. Army Medical Research and Materiel Command Grant DAMD17-1-8256, and by NIH/NCI grant R29 CA69421. Partial support was provided by the Huntsman Cancer Institute for the Utah Population Data Base. The Utah Cancer Registry was supported by NCI grant NO1 PC 67000.

In addition, the authors would like to thank Professor Andrei Y. Yakovlev for many helpful discussions.

REFERENCES

1. Aalen OO. Modeling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability* 1992;2:951-972.

2. Easton DF. The inherited component of cancer. *British Medical Bulliten* 1994;50;527-535.

3. Ford D. Easton DF. The genetics of breast and ovarian cancer. *Br. J. Cancer* 1995;72:805-812.

4. Newman B, Mu H, Butler LM, Millikan RC, Moorman PG, King MC. Frequency of breast cancer attributable to BRCA1 in a population-based series of American women *JAMA* 1998;279:915-921.

5. Anderson TI. Genetic heterogeneity in breast cancer susceptibility. *Acta Oncol.* 1996;35:407-410.

6. Vehmanen P, Friedman LS, Eerola H, Sarantaus L, Pyrh.onen S, Ponder BAJ, Muhonen T *et al.* A low proportion of BRCA2 mutations in Finnish breast cancer families. *Am J. Human Genet.* 1997;60:1050-1058.

7. Boucher KM, Kerber, RA. The shape of the hazard function for cancer incidence. *Math. and Computer Modeling* (in press).

8. Moolgavkar SH, Stevens RG , Lee JAH. Effect of age on incidence of breast cancer in females. *J National Cancer Institute* 1979;62:493-501

9. Khoury M, Flanders WD. Bias in using family history as a risk factor in case-control studies of disease. *Epidemiology* 1995;6:511-519.

10. Kerber RA. Method for calculating risk associated with family history of a disease. *Genetic Epidemiology* 1995;12:291-301.

11. Parmigiani G, Berry DA, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *A. J. Hum. Genet.* 1998;62:145-158

12. Keiding N. Independent delayed entry. In: Klein JP. and Goel PK., eds, *Survival Analysis: the State of the Art*, Kluwer, Boston-Dordrecht-London, 1992:309-326.

13. Etezadi-Amoli J, Ciampi A. Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function. *Biometrics* 1987;43:181-192.

14. Himmelblau DM. *Applied Nonlinear Programming* McGraw-Hill, Austin, 1972.

15. Malecot G. *Les Mathematiques de l'Hereditie* Masson, Paris, 1948.

16. Greenland, S, Robins, JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 1991;2:244-251.

17. Easton DF, Ford D, Bishop DT, Breast Cancer Linkage Consortium. Breast and ovarian cancer incidence in BRCA1-mutation carriers. *Am J Hum Genet* 1995;56:265-271.

Legends to figures

Figure 1. Distribution of the posterior probability that a subject carries BRCA1.

Figure 2. Distribution of the posterior probability that a subject carries BRCA2.

Figure 3. The distribution of loglog($FSIR$) before (A) and after (B) empirical Bayes adjustment.

Figure 4. Spline and life table estimates of the age-specific hazard for breast cancer, stratified by number of first degree relatives (A), posterior probability of BRCA1(B), posterior probability of BRCA2 (C), and empirically-Bayes adjusted FSIR. The high risk category contains about 8.5% of the subjects in each case.

Figure 5. Stratified spline-based estimates and 95% bootstrap confidencd bands for the age-specific hazard function for breast cancer. The categories are percentiles 0-75 (A), 75-90(B), 90-99.9 (C), and 99.9-100 (D) of the adjusted FSIR distribution. The scales are different, for better resolution.

Table 1. Number of subjects and range of the risk categories for the dichotomized familial risk variables. NFIRST refers to the number of first degree relatives with breast cancer. Pr(BRCA1) and Pr(BRCA2) refer to the posterior probability of carrying BRCA1 or BRCA2 from the model of Parmigiani, and FSIR refers to the familial standardized incidence ratio.

Risk Variable	Low Risk		High Risk	
	subjects	range	subjects	range
NFIRST	104680	0	9749	1-4
Pr(BRCA1)	104442	0-0.000452	9779	0.000452-0.96
Pr(BRCA2)	104440	0-0.000173	9781	0.000173-0.335
FSIR	104664	0.01-2.0	9765	2.0-6.1

Table 2. Stratification of FSIR for analysis with four categories, together with the number of cases per category.

Percentile of FSIR	Range	Subjects	Cases (% Cases)
¡ 75	0.01-1.2	85822	3279 (3.8%)
75-90	1.2-1.7	17165	951 (5.5%)
90-99.9	1.7-4.1	11328	845 (7.5%)
99.9-100	4.1-6.1	114	17 (14.9%)

Table 3. Likelihood ratio statistics estimates for models with posterior probabilities of BRCA1 and BRCA2 or their logarithms, as well as FSIR. see text for details. The chi-square value was computed using the likelihood ratio statistic.

Variable	χ_1^2
Pr(BRCA1)	8.52
Log(Pr(BRCA1)	44.94
Pr(BRCA2)	5.52
Log(Pr(BRCA2))	64.32
Loglog($FSIR$)	316.72

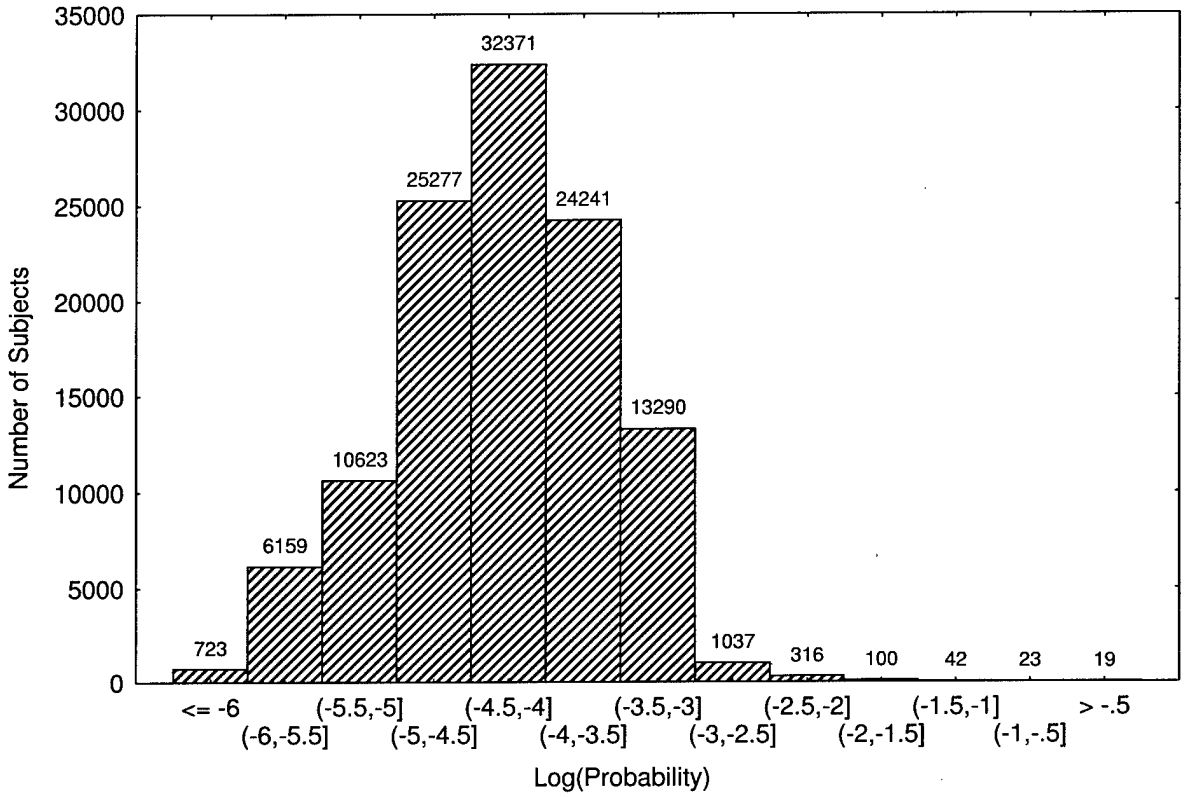

35000
30000
25000
20000
15000
10000
5000
0
Number of Subjects
723
6159
10623
25277
32371
24241
13290
1037
316
100
42
23
19
<= -6
(-6,-5.5]
(-5.5,-5]
(-5,-4.5]
(-4.5,-4]
(-4,-3.5]
(-3.5,-3]
(-3,-2.5]
(-2.5,-2]
(-2,-1.5]
(-1.5,-1]
(-1,-.5]
> -.5
Log(Probability)

FIGURE 1

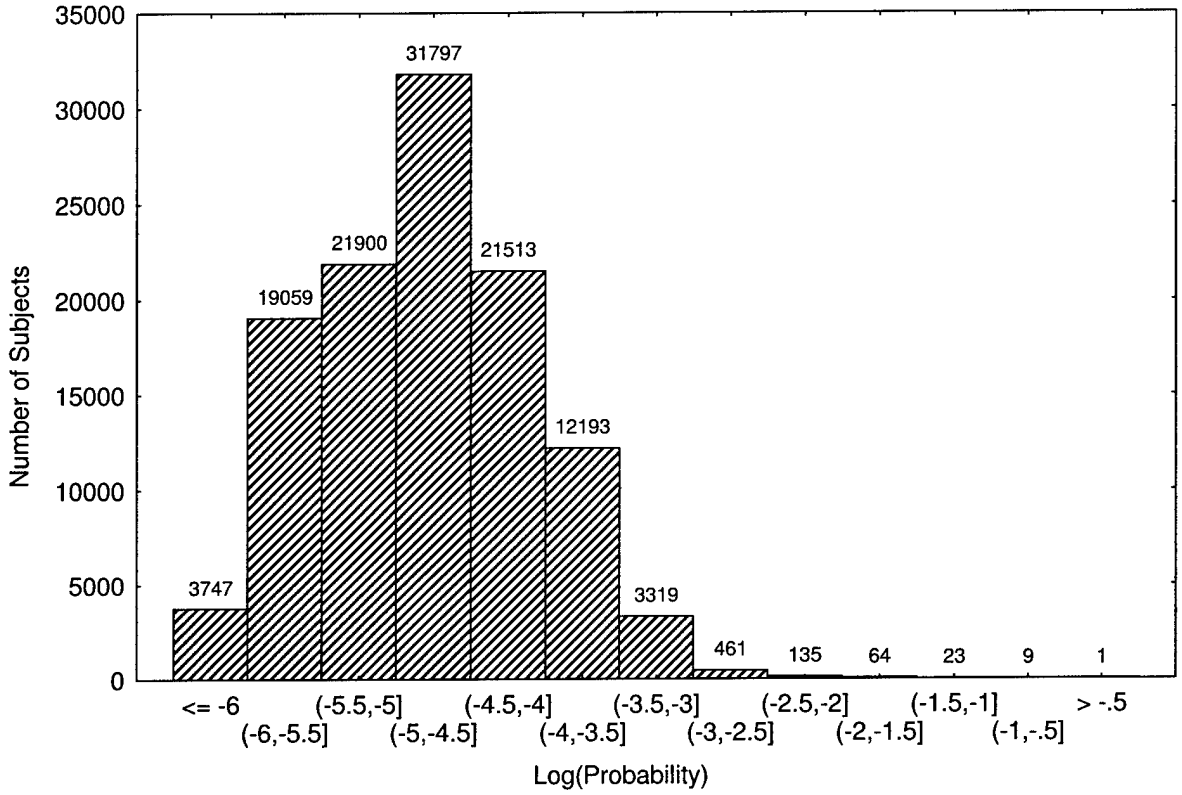

35000
30000
25000
20000
15000
10000
5000
0
Number of Subjects
3747
19059
21900
31797
21513
12193
3319
461
135
64
23
9
1
<= -6
(-6,-5.5]
(-5.5,-5]
(-5,-4.5]
(-4.5,-4]
(-4,-3.5]
(-3.5,-3]
(-3,-2.5]
(-2.5,-2]
(-2,-1.5]
(-1.5,-1]
(-1,-.5]
> -.5
Log(Probability)

FIGURE 2

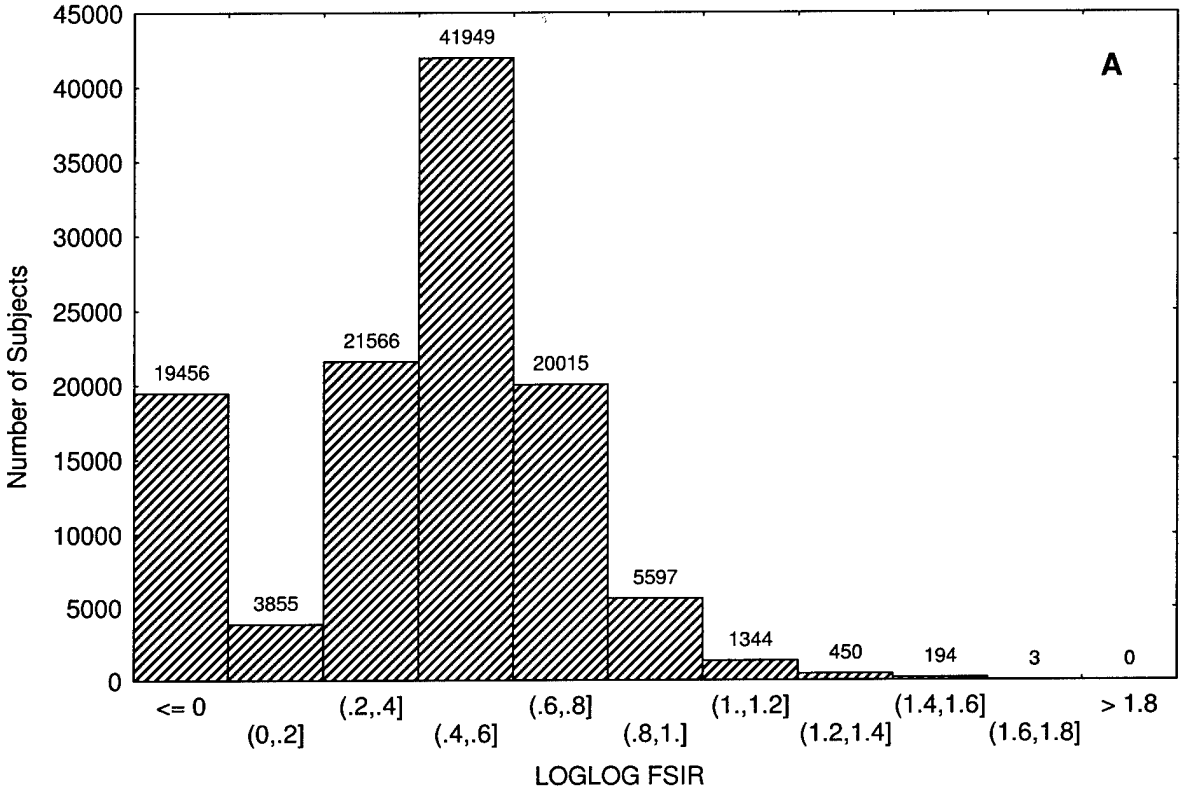

45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Number of Subjects
A
19456
3855
21566
41949
20015
5597
1344
450
194
3
0
<= 0
(0,.2]
(.2,.4]
(.4,.6]
(.6,.8]
(.8,1.]
(1.,1.2]
(1.2,1.4]
(1.4,1.6]
(1.6,1.8]
> 1.8
LOGLOG FSIR

FIGURE 3A

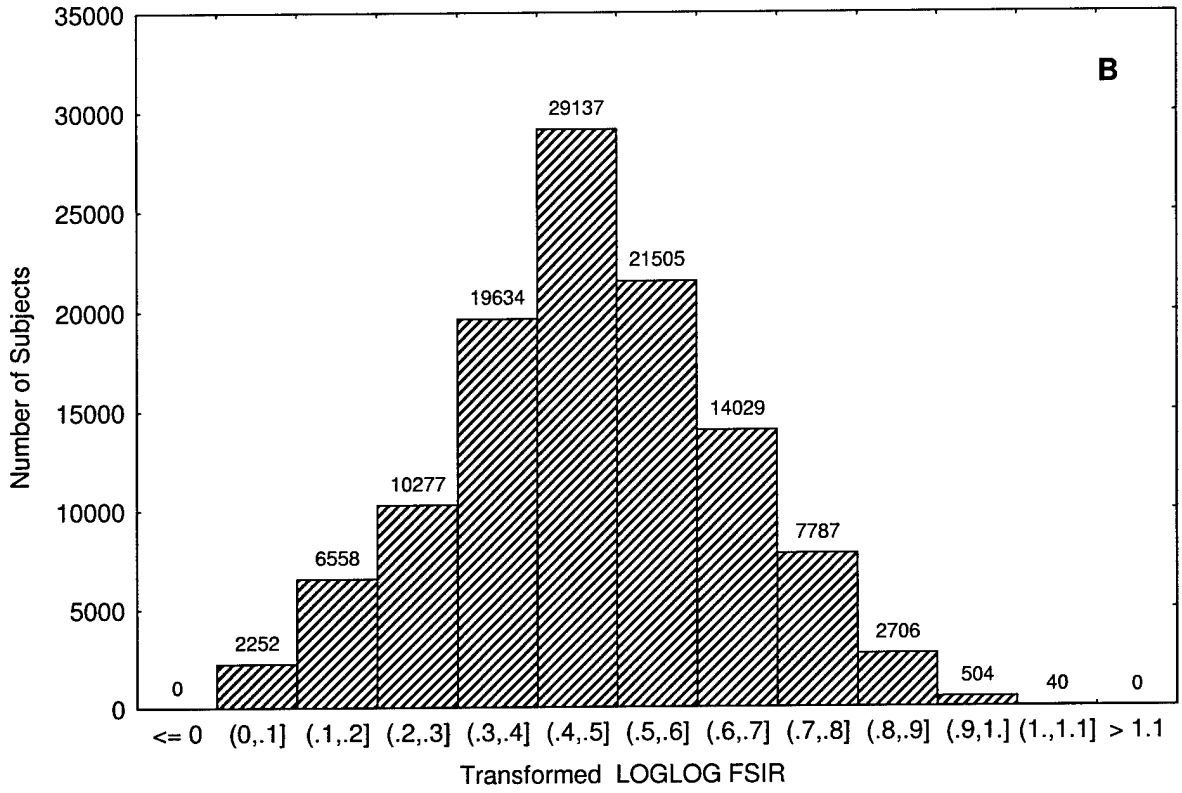

B
Number of Subjects
35000
30000
25000
20000
15000
10000
5000
0
0
2252
6558
10277
19634
29137
21505
14029
7787
2706
504
40
0
<= 0
(0,.1]
(.1,.2]
(.2,.3]
(.3,.4]
(.4,.5]
(.5,.6]
(.6,.7]
(.7,.8]
(.8,.9]
(.9,1.]
(1.,1.1]
> 1.1
Transformed LOGLOG FSIR

FIGURE 3B

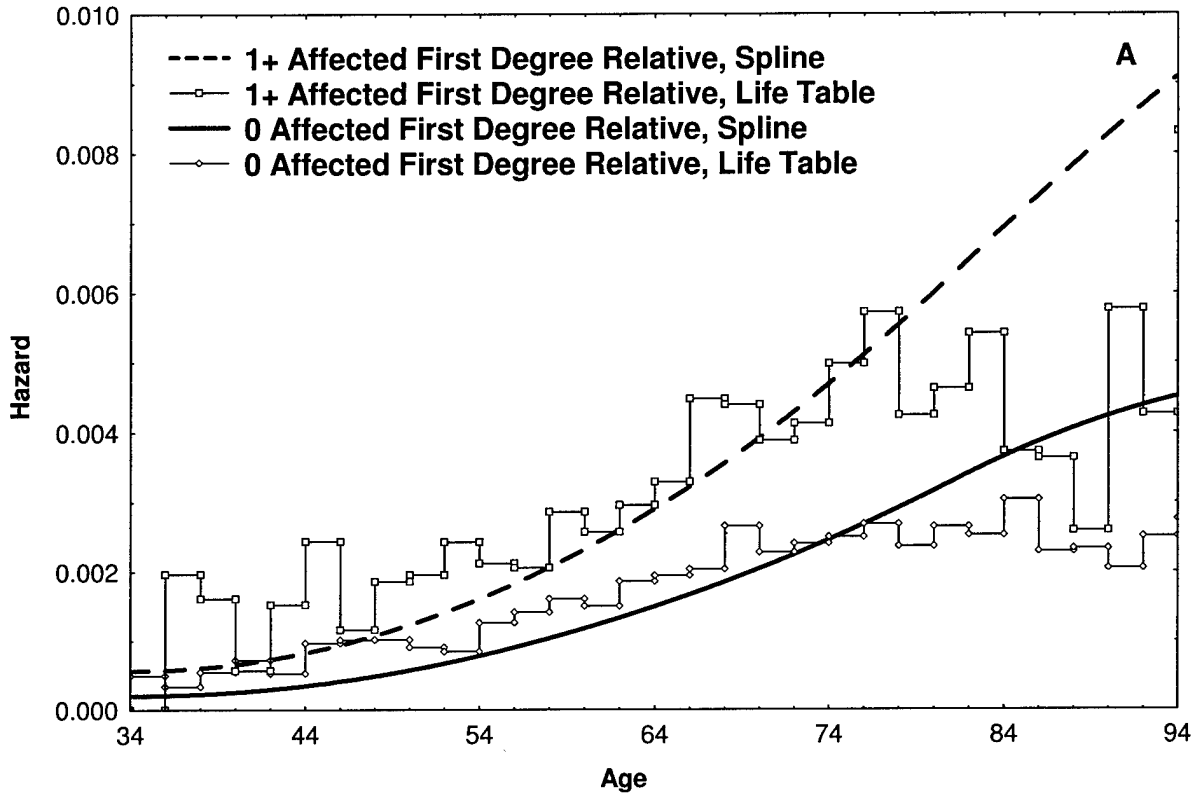

1+ Affected First Degree Relative, Spline
1+ Affected First Degree Relative, Life Table
0 Affected First Degree Relative, Spline
0 Affected First Degree Relative, Life Table
A
Hazard
0.010
0.008
0.006
0.004
0.002
0.000
34
44
54
64
74
84
94
Age

FIGURE 4A

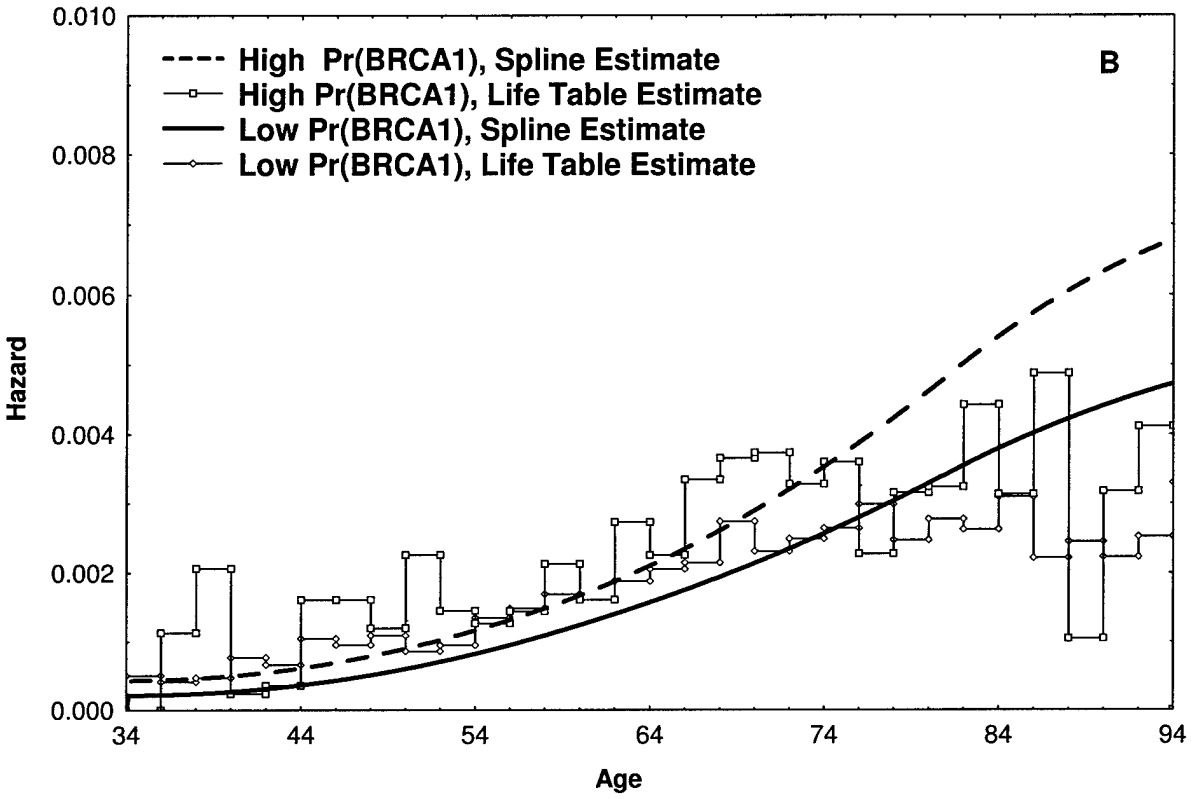

High Pr(BRCA1), Spline Estimate
High Pr(BRCA1), Life Table Estimate
Low Pr(BRCA1), Spline Estimate
Low Pr(BRCA1), Life Table Estimate
B
Hazard
0.010
0.008
0.006
0.004
0.002
0.000
34
44
54
64
74
84
94
Age

FIGURE 4B

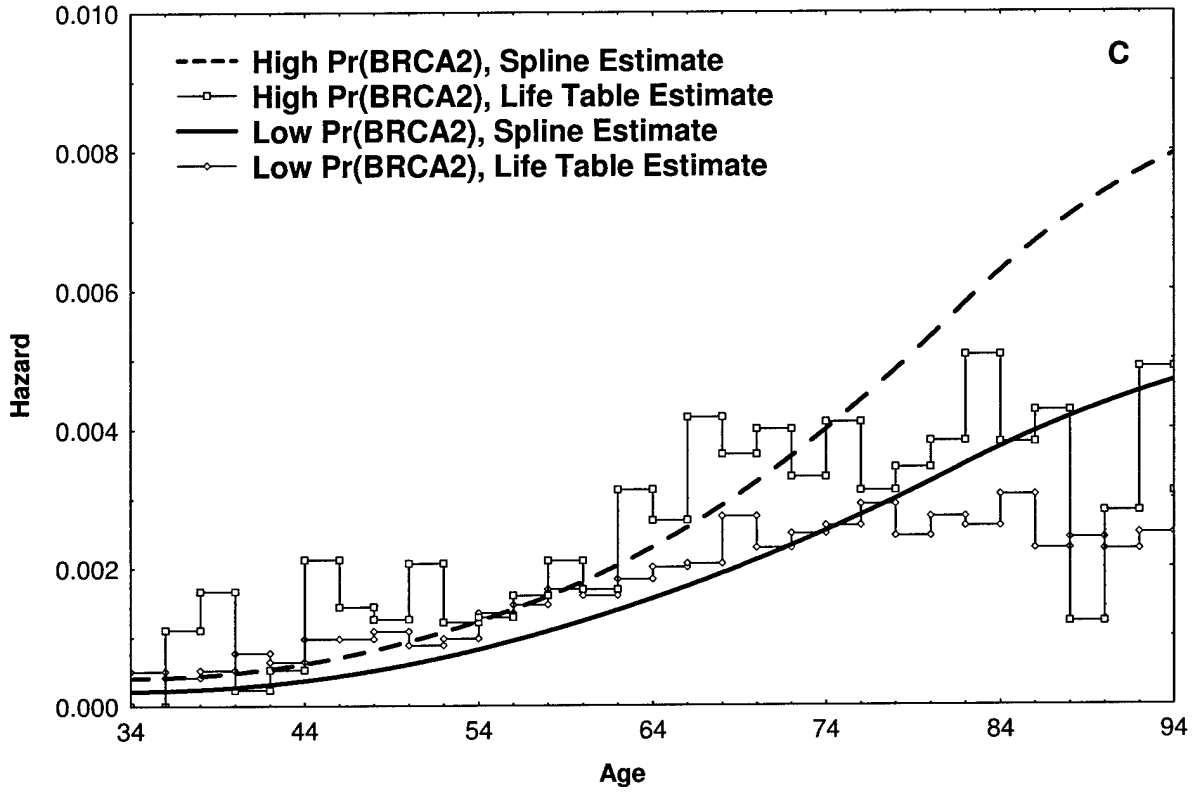

C
High Pr(BRCA2), Spline Estimate
High Pr(BRCA2), Life Table Estimate
Low Pr(BRCA2), Spline Estimate
Low Pr(BRCA2), Life Table Estimate
Hazard
0.010
0.008
0.006
0.004
0.002
0.000
34
44
54
64
74
84
94
Age

FIGURE 4C

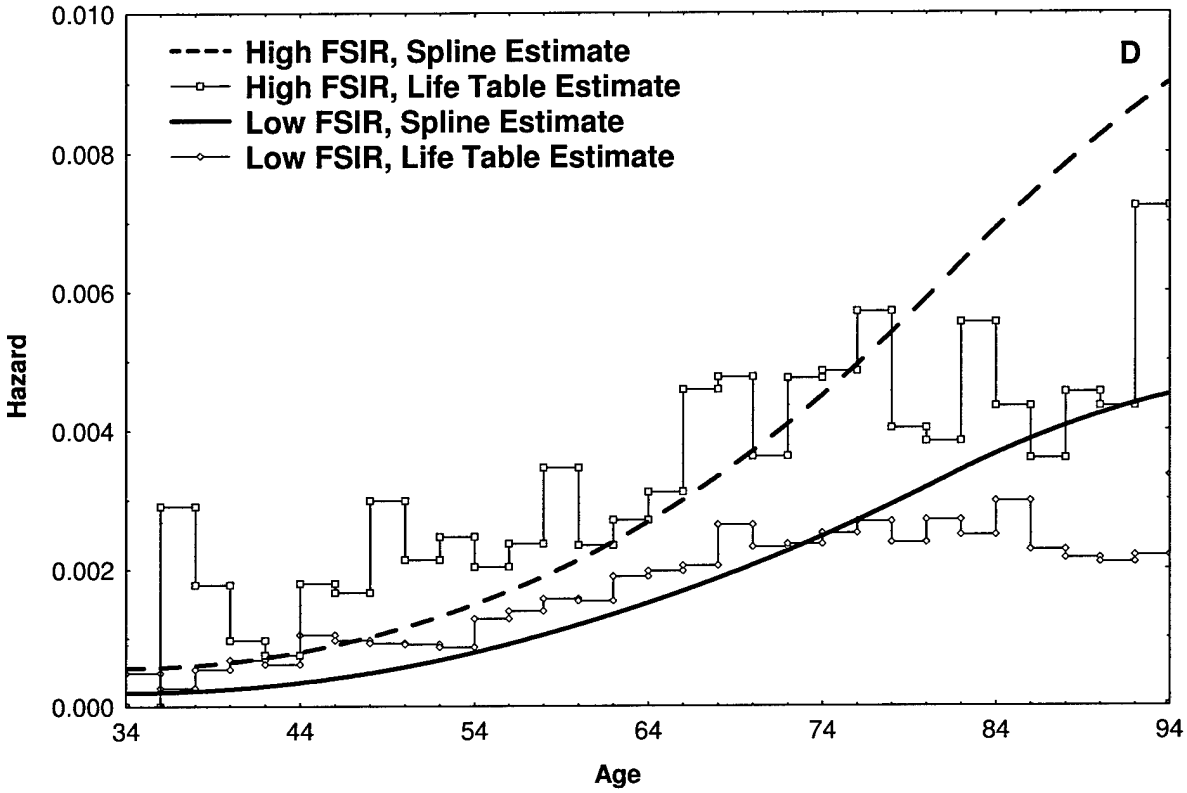

High FSIR, Spline Estimate
High FSIR, Life Table Estimate
Low FSIR, Spline Estimate
Low FSIR, Life Table Estimate
D
Hazard
0.010
0.008
0.006
0.004
0.002
0.000
34
44
54
64
74
84
94
Age

FIGURE 4D

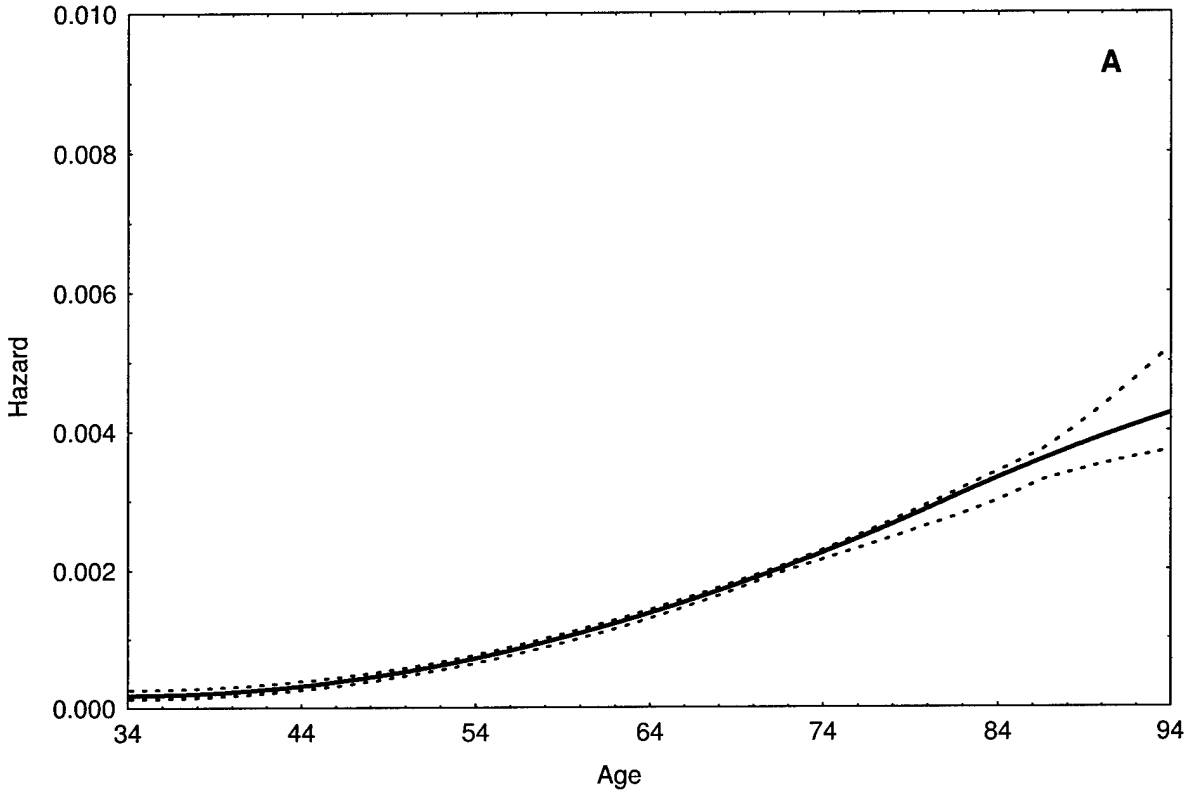


FIGURE 5A

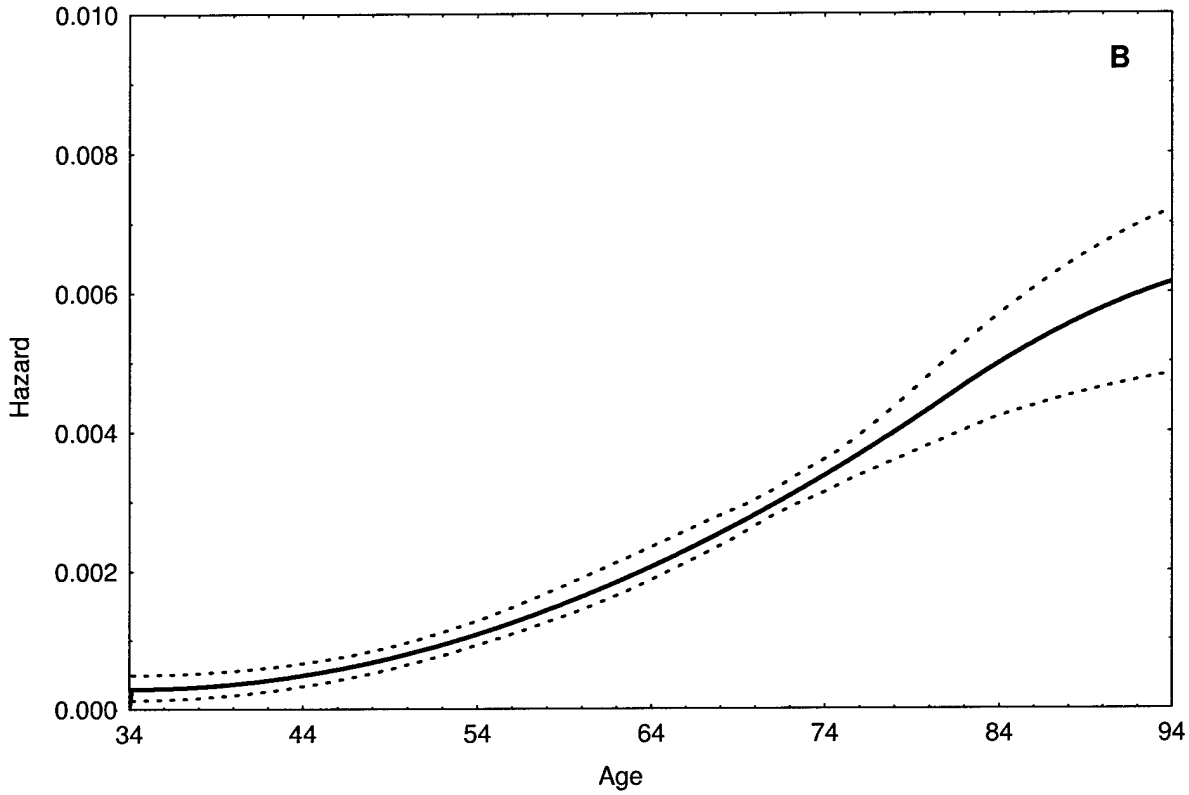


FIGURE 5B

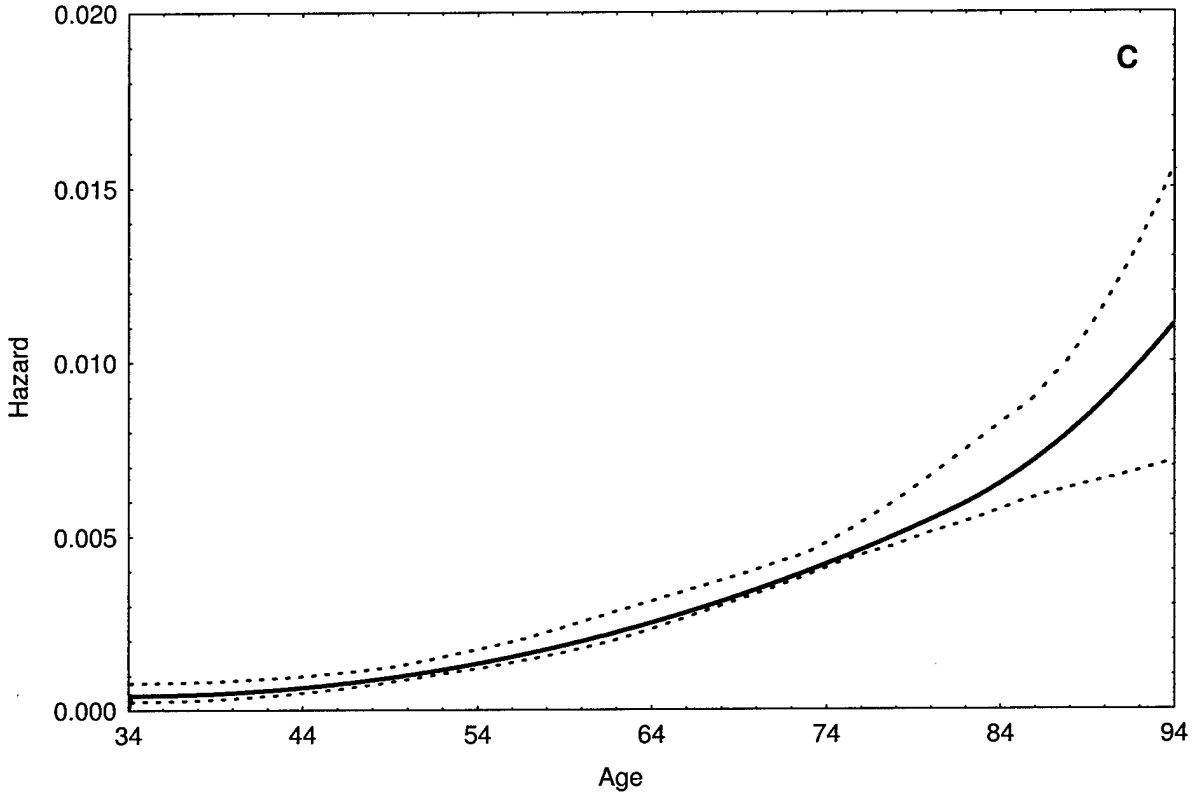


FIGURE 5C

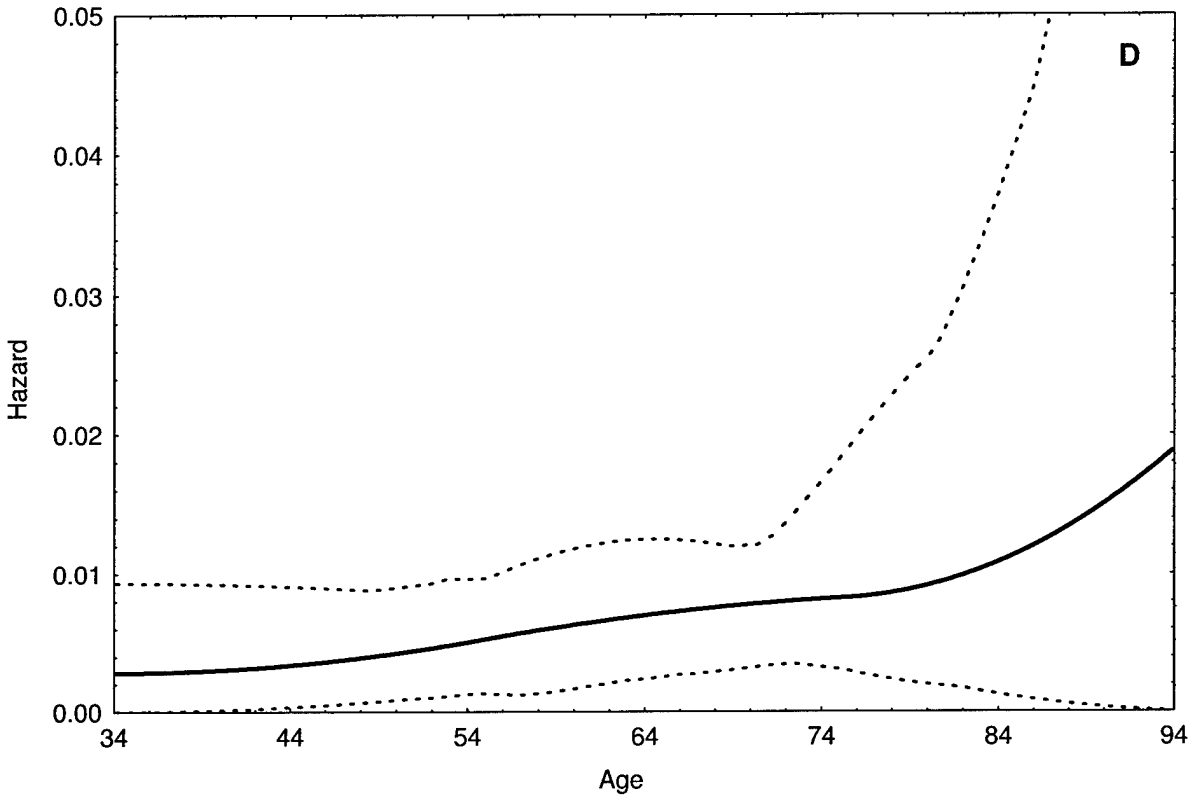


FIGURE 5D