STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:1837–1848 (DOI: 10.1002/sim.1428)

Age-conditional probabilities of developing cancer[‡]

Michael P. Fay[1,*,†], Ruth Pfeiffer[2], Kathleen A. Cronin[1], Chenxiong Le[3] and Eric J. Feuer[1]

[1] *National Cancer Institute, 6116 Executive Blvd, Suite 504, MSC 8317, Bethesda, Maryland 20892-8317, U.S.A.*
[2] *National Cancer Institute, 6120 Executive Blvd, EPS Room 8030, Bethesda, MD 20892-7335, U.S.A.*
[3] *Division of Biometrics III, CDER, FDA, HFD-725, Room N250, 9201 Corporate Blvd, Rockville, MD 20850, U.S.A.*

SUMMARY

We propose an estimator of the probability of developing a disease in a given age range, conditional on never having developed the disease prior to the beginning of the age range. Our estimator improves the one described by Wun, Merrill and Feuer (*Lifetime Data Analysis* 1998; **4**, 169–186) that is currently used by the U.S. National Cancer Institute for the SEER Cancer Statistics Review. Both estimators use cross-sectional disease rates and provide an interpretation of these rates in terms of the age-conditional probability of developing disease in a hypothetical cohort. The difficulty of this problem is that rates are not available per person-years alive and disease free, but only per person-years alive. Wun *et al.* used *ad hoc* methods to handle this problem which did not properly account for competing risks, did not provide a measure of variability, and only allowed age ranges using prespecified 5-year age intervals. Here we solve the problem under a unified competing risks framework, which allows the calculation of the age-conditional probabilities for any age range. We generalize gamma confidence intervals to apply to our new statistic. Although our new method provides estimates which are numerically similar to that of Wun *et al.*, this paper provides a comprehensive theoretical basis for estimation and inference about the age-conditional probability of developing a disease. Published in 2003 by John Wiley & Sons, Ltd.

KEY WORDS: competing risks; gamma confidence interval; hypothetical cohort; lifetime risk; surveillance; vital rates

1. INTRODUCTION

In this paper we use cross-sectional cancer rates and death rates to estimate lifetime and age-conditional probabilities of developing different types of cancer in a hypothetical cohort. If rates per person-years alive and cancer free are available then the estimation of these

[*] Correspondence to: Michael P. Fay, National Cancer Institute, 6116 Executive Boulevard, Suite 504, MSC 8317, Bethesda, Maryland 20892-8317, U.S.A.
[†] E-mail: faym@mail.nih.gov
[‡] This article is a US government work and is in the public domain in the U.S.A.

Received June 2002
Accepted September 2002

Published in 2003 by John Wiley & Sons, Ltd.

probabilities is a straightforward application of competing risk methodology. The difficulty is that many disease registries (including the National Cancer Institute's Surveillance, Epidemiology and End Results [SEER] cancer incidence data and National Center for Health Statistics [NCHS] mortality data that we use as our example) provide only rates per person-years alive. We show how to write the age-conditional probability of developing cancer as a function of the available rates, under a simple, standard assumption. In addition, we generalize the gamma confidence intervals developed for linear combinations of independent Poisson random variables [1], to apply to these more complex estimators.

Previous work on this problem is described in Wun *et al.* [2] and historical references may be found there. Wun *et al.* [2] did not fully account for competing risks in their model. Although they did use the theory of competing risks for some parts of their derivation (see equation (3) of Wun *et al.* [2]), some omissions were made in fully utilizing the competing risks framework. For example, in deriving the probability of developing cancer among the total population from the incidence rate, Wun *et al.* [2] used the formula for a failure time to a single event instead of the proper formula that accounts for competing risks (see equation (9) of Wun *et al.* [2]). This paper presents a new method for calculating the age-conditional probability of developing cancer which comprehensively accounts for competing risks. In Section 6 we compare our method with that of Wun *et al.* [2].

In Section 2 we review competing risk methods and in the process introduce our notation. In Section 3.1 we derive our estimator for the age-conditional probability of cancer. In Section 3.2 we provide methods to calculate confidence intervals. In Section 4 we apply the method to data examples. In Section 5 we explore the properties of our confidence interval estimator through simulation. A concluding discussion is presented in Section 6.

2. NOTATION AND REVIEW OF COMPETING RISK METHODS

Consider first the standard competing risk problem (see, for example, Kalbfleisch and Prentice [3]). We observe the time until one of several events, T, and an indicator of the type of event that occurred, J. In this paper, T is a random variable denoting the age at death and J has one of two values, $J=\mathrm{d}$ means death from the event of interest (for example, breast cancer), and $J=\mathrm{o}$ means death from other causes. For ease of exposition, we use the term 'cancer' to denote the event of interest. The cause specific hazard function for $J=j$ is

$$\lambda_j(a)=\lim_{\varepsilon\to 0^+}\frac{\Pr[a\leqslant T<a+\varepsilon, J=j|T\geqslant a]}{\varepsilon}$$

Thus $\lambda_\mathrm{d}(a)$ is the rate of cancer deaths per person-years alive at age a, and $\lambda_\mathrm{o}(a)$ is the rate of other (that is, non-cancer) deaths per person-years alive at age a. The overall failure rate at age a is $\lambda(a)=\lambda_\mathrm{d}(a)+\lambda_\mathrm{o}(a)$, and the overall survival function is $S(a)=\Pr[T>a]=\exp(-\int_0^a \lambda(u)\,\mathrm{d}u)$. The probability of dying from cause j in the age interval $[x,y)$ given survival until just prior to x is

$$\Pr[x\leqslant T<y, J=j|T\geqslant x]=\frac{\int_x^y \lambda_j(u)S(u-)\,\mathrm{d}u}{S(x-)}$$

where $S(a-)=\lim_{\varepsilon\to 0} S(a-\varepsilon)$.

We also consider the statistically identical competing risks problem where T^* is the age at either first cancer or death before first cancer, and J^* is the indicator with $J^* = \mathrm{c}$ denoting that T^* is the age at first cancer and $J^* = \mathrm{o}$ denoting that T^* is the age at death if death occurs before the first cancer. The cause specific hazard functions are: $\lambda_\mathrm{c}^*(a)$, the rate of first cancer per person-years alive *and cancer free* at age a, and $\lambda_\mathrm{o}^*(a)$, the rate of deaths per person-years alive *and cancer free* at age a. Then, similar to above, the probability of getting a first cancer in the age interval $[x, y)$ given alive and cancer free until just prior to x is

$$A(x,y) = \Pr[x \leqslant T^* < y, J^* = \mathrm{c} | T^* \geqslant x] = \frac{\int_x^y \lambda_\mathrm{c}^*(u) S^*(u-)\,\mathrm{d}u}{S^*(x-)} \tag{1}$$

where $S^*(a) = \exp\{-\int_0^a \lambda^*(u)\,\mathrm{d}u\}$ and $\lambda^*(a) = \lambda_\mathrm{c}^*(a) + \lambda_\mathrm{o}^*(a)$.

3. AGE CONDITIONAL PROBABILITIES OF DEVELOPING CANCER ESTIMATED FROM CANCER REGISTRIES

3.1. The estimator

We wish to estimate $A(x, y)$ as given in equation (1), but we cannot directly obtain estimators of either $\lambda_\mathrm{c}^*(a)$ or $\lambda^*(a)$, the rates of cancer and total deaths, respectively, per person-years alive and cancer free. However, we can directly estimate the following rates per person-years alive at age a: $\lambda_\mathrm{c}(a)$, the rate of first cancer incidence; $\lambda_\mathrm{d}(a)$, the rate of cancer deaths; $\lambda_\mathrm{o}(a)$, the rate of other (that is, non-cancer) deaths. We assume that the rate of non-cancer deaths is the same for all people regardless of whether or not they have had a cancer, so that $\lambda_\mathrm{o}^*(a) = \lambda_\mathrm{o}(a)$. After making this assumption, we show in the following that we can rewrite $A(x, y)$ in terms of the functions $\lambda_\mathrm{c}(\cdot), \lambda_\mathrm{d}(\cdot)$ and $\lambda_\mathrm{o}(\cdot)$.

First consider the numerator of equation (1). Rewrite $\lambda_\mathrm{c}(a)$ as

$$\begin{aligned}\lambda_\mathrm{c}(a) &= \lim_{\varepsilon\to 0^+} \frac{\Pr[a \leqslant T^* < a+\varepsilon, J^* = \mathrm{c} | T \geqslant a]}{\varepsilon} = \lim_{\varepsilon\to 0^+} \frac{\Pr[a \leqslant T^* < a+\varepsilon, J^* = \mathrm{c} \text{ and } T \geqslant a]}{\varepsilon \Pr[T \geqslant a]} \\ &= \lim_{\varepsilon\to 0^+} \frac{\Pr[a \leqslant T^* < a+\varepsilon, J^* = \mathrm{c}]}{\varepsilon \Pr[T \geqslant a]} = \frac{\lambda_\mathrm{c}^*(a) S^*(a-)}{S(a-)} \end{aligned} \tag{2}$$

Using this equation we write the numerator of equation (1) as $\int_x^y \lambda_\mathrm{c}(u) S(u-)\,\mathrm{d}u$.

Rewrite the denominator, $S^*(a) = S_\mathrm{c}^*(a) S_\mathrm{o}^*(a)$, where $S_j^*(a) = \exp(-\int_0^a \lambda_j^*(u)\,\mathrm{d}u)$ for $j = \mathrm{c}, \mathrm{o}$. Note that $S_j^*(a)$ does not have a survival function interpretation (see Kalbfleisch and Prentice, reference [3, p. 168]). Because we have assumed that $\lambda_\mathrm{o}^*(a) = \lambda_\mathrm{o}(a)$, we write $S_\mathrm{o}^*(a) = S_\mathrm{o}(a) = \exp(-\int_0^a \lambda_\mathrm{o}(u)\,\mathrm{d}u)$, and the only outstanding problem is finding an estimator of $S_\mathrm{c}^*(a)$. In the definition of $S_\mathrm{c}^*(a)$, we rewrite the expression for $\lambda_\mathrm{c}^*(a)$ using equation (2), and we obtain the recursive equation

$$S_\mathrm{c}^*(a) = \exp\left\{-\int_0^a \frac{\lambda_\mathrm{c}(u) S(u-)}{S_\mathrm{o}(u-) S_\mathrm{c}^*(u-)}\,\mathrm{d}u\right\} \tag{3}$$

To solve this recursive equation, first let $S(t) = S_d(t)S_o(t)$, where $S_j(t) = \exp(-\int_0^t \lambda_j(u)\,du)$ for $j = d, o$. Using the assumption that $\lambda_o(a) = \lambda_o^*(a)$, equation (3) becomes

$$S_c^*(t) = \exp\left\{ -\int_0^t \frac{\lambda_c(u)S_d(u-)}{S_c^*(u-)}\,du \right\}$$

Take log of both sides, then differentiate with respect to t to get

$$\frac{\frac{dS_c^*(t)}{dt}}{S_c^*(t)} = -\frac{\lambda_c(t)S_d(t-)}{S_c^*(t-)}$$

If T^* is a continuous random variable then $S_c^*(t) = S_c^*(t-)$ and $dS_c^*(t)/dt = -\lambda_c(t)S_d(t-)$. Now integrate to obtain $S_c^*(a) - S_c^*(0) = -\int_0^a \lambda_c(t)S_d(t-)\,dt$ and $S_c^*(0) = 1$, so that $S_c^*(a) = 1 - \int_0^a \lambda_c(u)S_d(u-)\,du$. Thus, under the assumption $\lambda_o^*(a) = \lambda_o(a)$, $A(x, y)$ can be expressed as

$$A(x, y) = \frac{\int_x^y \lambda_c(u)S(u-)\,du}{S_o(x-)\{1 - \int_0^x \lambda_c(u)S_d(u-)\,du\}} \tag{4}$$

To obtain an estimate of $A(x, y)$ using SEER incidence data and NCHS mortality data, we first divide the possible ages into $k+1$ intervals, $[a_i, a_{i+1})$ where $0 = a_0 < a_1 < \cdots < a_k < a_{k+1} = \infty$, and choose a calendar interval, $[t_1, t_2)$. We observe the number of first cancer incident cases (c_i), cancer deaths (d_i), and other deaths (o_i), occurring at ages in the interval $[a_i, a_{i+1})$ during the calendar time $[t_1, t_2)$, for $i = 0, \ldots, k$. Although the cancer incident cases and the deaths often come from the same population (see Table I), this is not necessary (see Table III). We also observe $n_i^{(j)}$, which is $(t_2 - t_1)$ times the estimated number of people from the same population associated with event j (where $j = c, d$, or o) with ages in $[a_i, a_{i+1})$ at the midpoint, $(t_1 + t_2)/2$, of the interval $[t_1, t_2)$, for $i = 0, \ldots, k$. If $t_2 - t_1 = 1$, $n_i^{(j)}$ corresponds to the midyear population with ages in $[a_i, a_{i+1})$.

We assume that the observed counts c_i, d_i, o_i are Poisson and the midinterval populations are fixed constants. For a motivation and discussion of this assumption see Brillinger [4] with discussion (see especially the discussion by Keiding). Assuming constant rates within age intervals, we estimate rates for ages $a \in [a_i, a_{i+1})$ by $\hat{\lambda}_c(a) = c_i/n_i^{(c)}$, $\hat{\lambda}_d(a) = d_i/n_i^{(d)}$, and $\hat{\lambda}_o(a) = o_i/n_i^{(o)} = \hat{\lambda}_o^*(a)$. These estimators replace their associated functions in equation (4) to obtain our estimator of $A(x, y)$. In Appendix A we show the estimator using summation notation.

Because we are using cross-sectional data from finite populations to estimate hazard rates for a hypothetical cohort, these estimates may produce hazards that cannot possibly describe a real cohort. There are two types of these ‘impossible’ hazards. If no one in the oldest age group dies (that is, $d_k = 0$ and $o_k = 0$), then the resulting hazards describe an impossible cohort where the probability of living forever is non-zero. Another impossible cohort would result if the probability of dying of cancer by any age a is greater than the probability of getting cancer by that same age a (this is equivalent to $\int_0^a \lambda_d(u)\,du > \int_0^a \lambda_c(u)\,du$). These impossible cohorts would rarely occur in large populations.

Table I. Raw data.

Age, years	Female breast cancer, invasive only 11 SEER registries, 1996–1998				Acute lymphocytic leukaemia, both sexes 9 SEER registries, 1990			
	c_i	o_i	d_i	$n_i^{(c)}=n_i^{(o)}=n_i^{(d)}$	c_i	o_i	d_i	$n_i^{(c)}=n_i^{(o)}=n_i^{(d)}$
$[0,5)$	0	5893	0	4052953	97	4096	10	1817956
$[5,10)$	0	561	0	4032790	61	335	12	1724041
$[10,15)$	1	627	1	3784789	24	360	12	1629304
$[15,20)$	9	1367	0	3810986	20	1375	13	1614939
$[20,25)$	43	1541	6	3675646	7	1898	14	1780348
$[25,30)$	335	2029	35	4138795	8	2399	4	2066277
$[30,35)$	1116	3012	173	4575728	10	3266	8	2153289
$[35,40)$	2670	4531	425	4831799	8	3884	9	1984257
$[40,45)$	5183	6234	765	4578168	14	4423	6	1776224
$[45,50)$	7392	8065	1152	3906260	9	4716	7	1349233
$[50,55)$	8012	9976	1427	3054146	11	5708	8	1064862
$[55,60)$	7341	12424	1411	2353577	9	8144	3	956807
$[60,65)$	7010	16957	1436	1981443	9	12837	4	958029
$[65,70)$	7651	25818	1668	1988371	4	18117	6	901014
$[70,75)$	8060	39434	1920	1838556	14	21592	10	712642
$[75,80)$	7146	51697	1800	1541002	11	25109	10	535934
$[80,85)$	4754	62624	1533	1083867	7	24924	7	340699
$[85,90)$	2574	63851	1081	629172	6	21139	5	183481
$[90,95)$	952	48324	531	299128	0	13316	0	71081
$[95,\infty)$	273	26926	232	114178	1	6781	1	23807

3.2. *Confidence limits for $A(x,y)$*

In this section we modify the gamma confidence intervals, developed for linear combinations of independent Poisson random variables by Fay and Feuer [1], to create confidence intervals for $A(x,y)$. First, we put all the Poisson counts into one $(3K+3)\times 1$ vector

$$\boldsymbol{z}=[z_1,z_2,\ldots,z_{3k+3}]'=[c_0,c_1,\ldots,c_k,d_0,d_1,\ldots,d_k,o_0,o_1,\ldots,o_k]'$$

Associated with each z_i is a random variable Z_i which we assume has a Poisson distribution with mean μ_i. Let $\boldsymbol{\mu}=[\mu_1,\mu_2,\ldots,\mu_{3k+3}]'$. In the previous notation $\boldsymbol{\mu}=[\lambda_c(a_0)n_0^{(c)},\lambda_c(a_1)n_1^{(c)},\ldots,\lambda_o(a_k)n_k^{(o)}]'$. Emphasizing the dependence of $A(x,y)$ on $\boldsymbol{\mu}$, we write $A(x,y)=A(x,y,\boldsymbol{\mu})$. Using this notation, our estimator is $A(x,y,\boldsymbol{z})$. For ease of exposition we write $A(x,y,\boldsymbol{z})=A(\boldsymbol{z})$ and $A(x,y,\boldsymbol{\mu})=A(\boldsymbol{\mu})$, suppressing dependence on x and y. Using a Taylor series expansion

$$A(\boldsymbol{z})\approx A(\boldsymbol{\mu})+\left[\frac{\partial A(\boldsymbol{t})}{\partial \boldsymbol{t}}\right]_{\boldsymbol{t}=\boldsymbol{\mu}}(\boldsymbol{z}-\boldsymbol{\mu}) \tag{5}$$

and

$$\operatorname{var}(A(\boldsymbol{Z}))\approx\left[\frac{\partial A(t)}{\partial t}\right]_{t=\boldsymbol{z}}\operatorname{diag}(\boldsymbol{z})\left[\frac{\partial A(t)}{\partial t}\right]_{t=\boldsymbol{z}}'$$

where $\operatorname{diag}(\boldsymbol{z})$ is a diagonal matrix with the values of $\boldsymbol{z}$ on the diagonal, representing an estimate of $\operatorname{var}(\boldsymbol{Z})$.

Alternatively, numerical derivatives can be used. Letting

$$\Delta A(\mathbf{z}) = \begin{bmatrix} A(x, y, \mathbf{z}_{(+1)}) - A(x, y, \mathbf{z}) \\ A(x, y, \mathbf{z}_{(+2)}) - A(x, y, \mathbf{z}) \\ \vdots \\ A(x, y, \mathbf{z}_{(+[3K+3])}) - A(x, y, \mathbf{z}) \end{bmatrix}$$

with $\mathbf{z}_{(+\ell)} = [z_1, \ldots, z_{\ell-1}, 1 + z_\ell, z_{\ell+1}, \ldots, z_{3K+3}]'$, leads to our variance estimate

$$V(\mathbf{z}) = \{\Delta A(\mathbf{z})\}\, \mathrm{diag}(\mathbf{z}) \{\Delta A(\mathbf{z})\}'$$

Our generalization of the gamma intervals [1] is to use the Taylor expansion as the linear combination of independent Poisson random variables. The only complication is that the weights may be negative and depend on the Poisson values. This complication does not effect the lower confidence limit, though; the $100(1-\alpha)$ per cent lower confidence limit is given by $L = G^{-1}_{\gamma,\beta}(\alpha/2)$ where $G^{-1}_{\gamma,\beta}(p)$ is the pth quantile of the gamma distribution with parameters $\gamma = \frac{A(\mathbf{z})^2}{V(\mathbf{z})}$ and $\beta = \frac{V(\mathbf{z})}{A(\mathbf{z})}$ (that is, with mean $A(\mathbf{z})$ and variance $V(\mathbf{z})$). However, for the upper limit the method has to be altered. When finding the maximum discrete increase in $A(\mathbf{z})$, it is possible that this may occur with a decrease in one of the Poisson values. Let

$$\mathbf{z}_{(-\ell)} = [z_1, \ldots, z_{\ell-1}, \max(0, -1 + z_\ell), z_{\ell+1}, \ldots, z_{3K+3}]'$$

Define $\mathbf{z}_{(M)}$ to be the vector value of either $\mathbf{z}_{(+\ell)}$ or $\mathbf{z}_{(-\ell)}$ for $\ell = 1, \ldots, 3K+3$ such that $A(\mathbf{z}_{(M)})$ is maximized. Then the upper confidence limit is $U = G^{-1}_{\gamma_M,\beta_M}(1 - \frac{\alpha}{2})$ where $\gamma_M = \frac{A(\mathbf{z}_{(M)})^2}{V(\mathbf{z}_{(M)})}$ and $\beta_M = \frac{V(\mathbf{z}_{(M)})}{A(\mathbf{z}_{(M)})}$. Note that if we let the population and the mean $\boldsymbol{\mu}$ get larger by the same constant, say N, then the generalized gamma intervals approach the usual delta method intervals (see, for example, Lehmann [5]) as $N \to \infty$ (see Appendix B). For small $\boldsymbol{\mu}$ these generalized gamma intervals perform better and are calculated straightforwardly even when some $z_i = 0$, while the delta method requires modification whenever some $z_i = 0$ in order to prevent estimates of zero variance. For the delta method, the variances corresponding to the elements $\mathbf{z}$ are estimated by replacing elements with $z_i = 0$ with 0.5.

4. EXAMPLES

Our examples use SEER cancer incidence data and NCHS mortality data associated with the corresponding SEER catchment areas (see Ries *et al.* [6]). We calculate our statistics for two types of cancer, invasive female breast cancer, one of the more common cancers, for all races from the expanded 11 SEER registries from $t_1 = 1$ January 1996 until $t_2 = 31$ December 1998, and acute lymphocytic leukaemia (ALL) for all races from the nine SEER registries active during $t_1 = 1$ January 1990 until $t_2 = 31$ December 1990. ALL was chosen because it is primarily a childhood cancer (see Table I) and provides an example which has high rates at young ages unlike many cancer sites, for example, breast cancer, which have increasing incidence for older age groups.

The raw data are listed in Table I, and $A(x, y, \mathbf{z})$ with the associated 95 per cent confidence intervals for different values of x and y are listed in Table II. As expected from Appendix

Table II. Estimated per cent developing cancer by age y, given no cancer before age x (with 95 per cent confidence intervals).

x	y	Female breast cancer invasive only 11 SEER registries, 1996–1998			Acute lymphocytic leukaemia, both sexes 9 SEER registries, 1990		
		$100 \times A(x,y,z)$	Gamma method	Delta method	$100 \times A(x,y,z)$	Gamma method	Delta method
0	30	0.0470	(0.0424, 0.0519)	(0.0423, 0.0517)	0.0612	(0.0533, 0.0699)	(0.0530, 0.0693)
0	50	1.8995	(1.8708, 1.9286)	(1.8707, 1.9284)	0.0722	(0.0637, 0.0817)	(0.0634, 0.0811)
0	70	7.7861	(7.7130, 7.8598)	(7.7128, 7.8594)	0.0867	(0.0769, 0.0976)	(0.0766, 0.0969)
0	∞	13.3198	(13.2170, 13.4235)	(13.2168, 13.4228)	0.1088	(0.0968, 0.1227)	(0.0964, 0.1213)
30	50	1.8817	(1.8529, 1.9108)	(1.8527, 1.9106)	0.0114	(0.0081, 0.0155)	(0.0078, 0.0149)
30	70	7.8609	(7.7868, 7.9355)	(7.7866, 7.9351)	0.0263	(0.0205, 0.0333)	(0.0201, 0.0325)
30	∞	13.4816	(13.3773, 13.5868)	(13.3771, 13.5861)	0.0491	(0.0399, 0.0602)	(0.0394, 0.0587)
50	70	6.2505	(6.1793, 6.3224)	(6.1791, 6.3220)	0.0157	(0.0108, 0.0219)	(0.0103, 0.0210)
50	∞	12.1264	(12.0217, 12.2320)	(12.0214, 12.2313)	0.0395	(0.0307, 0.0506)	(0.0301, 0.0490)
70	∞	7.3149	(7.2202, 7.4109)	(7.2199, 7.4100)	0.0302	(0.0213, 0.0422)	(0.0204, 0.0401)

B, the delta method confidence intervals are very similar to the gamma method confidence intervals for these data. We also calculated the gamma confidence intervals using the exact derivatives ($\partial A(t)/\partial t$) instead of the numerical ones ($\Delta A(z)$), and these intervals (not shown) give essentially the same results as the gamma intervals listed in Table II (the values in terms of probabilities are equal up to five significant digits).

From Table II, the estimated probability of developing invasive breast cancer in one's lifetime is 0.1332 while the probability of developing breast cancer given alive and cancer free at 30 is 0.1348. It seems contradictory that by surviving from age 0 to age 30 without dying or getting breast cancer, a woman actually increases the probability of getting breast cancer in the remainder of her life. To gain more insight into this situation consider an example of a birth cohort of 100 females and assume that 12 will develop breast cancer over their life time. If by age 5 two of the girls have died of other causes and none have yet developed breast cancer, the risk of developing breast cancer after age 5 is $12/98(>12/100)$ in this cohort.

5. SIMULATIONS

We tested the coverage probabilities of our method in three situations. For the first two situations (female breast cancer and ALL) we assumed that the rates were exactly equal to the rates derived from Table I except we added 0.5 to zero counts. Then we simulated 10000 data sets assuming independent Poisson distributions with means equal to those counts (with 0.5 added to zeros). For the third situation, we checked our method for extremely low counts; we used incidence rates of eye and orbit cancer in the nine SEER areas in 1990 (after adding 0.5 to the zero value), and rates of eye and orbit cancer deaths and other deaths from the entire U.S. in 1990, and simulated these rates applied to the Vietnamese population in the nine SEER areas in 1990. The raw data are presented in Table III. Thus for example the expected value for c_1 is $5914 \times (28/1817468) = 0.0911$, and we have many expected count values that are much less than 1. Then we simulated 10000 data sets assuming

Table III. Raw data, eye and orbit cancer, both sexes, 1990.

Age, years	SEER 9 All races		Total U.S. All races			SEER 9 Vietnamese
	c_i	$n_i^{(c)}$	o_i	d_i	$n_i^{(o)} = n_i^{(d)}$	n_i
$[0, 5)$	28	1817468	45269	13	18852851	5914
$[5, 10)$	2	1723903	3992	3	18061843	6360
$[10, 15)$	1	1630063	4440	1	17198108	6249
$[15, 20)$	0	1613257	15710	1	17764520	8555
$[20, 25)$	1	1777565	21020	2	19134952	7627
$[25, 30)$	1	2064309	26578	1	21235575	7270
$[30, 35)$	7	2153703	33507	5	21912156	6895
$[35, 40)$	5	1986987	39089	4	19982168	6546
$[40, 45)$	10	1776946	44466	3	17794548	4516
$[45, 50)$	14	1349043	51850	6	13823785	3008
$[50, 55)$	15	1064803	66743	9	11369647	2111
$[55, 60)$	14	956860	97852	13	10474089	1442
$[60, 65)$	19	958022	154800	31	10619134	1047
$[65, 70)$	29	901564	217299	34	10076737	724
$[70, 75)$	35	713026	260584	32	8022791	591
$[75, 80)$	18	536271	301073	41	6146687	345
$[80, 85)$	8	340946	300298	27	3935220	180
$[85, \infty)$	4	278600	463076	29	3059585	82

independent Poisson distributions with means equal to those expected counts. We calculated a 95 per cent confidence interval for each simulation. In Table IV we list both E_L, the percentage of the lower confidence limits that are greater than the true value and E_U, the percentage of the upper confidence limits that are less than the true value, where the *true value* refers to the estimator calculated from the counts with 0.5 added to the zeros.

The situations with larger counts give better error rates, and the third situation with extremely low expected counts gives error rates that are very conservative. In each of the three situations the gamma intervals have error rates closer to the nominal 2.5 per cent than the delta method based confidence intervals, although there is essentially no difference in the first case. For ALL the asymmetric gamma confidence intervals produce more central confidence intervals (that is, the tails of the errors are more nearly equal) than the symmetric delta confidence intervals. For the eye and orbit situation, both methods perform very conservatively, but the gamma method is generally less conservative. In addition all the lower delta method confidence limits were less than 0, and most of the upper limits were greater than the gamma method upper limits. Thus, in all situations the gamma method performed better than the delta method.

6. DISCUSSION

In this paper we derived a new estimator of $A(x, y)$, the probability of developing a first time cancer during the age interval $[x, y)$, conditioned on being alive and cancer free at age

Table IV. Simulated error rates for 95 per cent confidence limits (ideal error rates are 2.5 per cent); 10000 simulations for each cancer/age range combination.

x	y	Simulated using female breast cancer invasive only, 11 SEER registries, 1996–1998				Simulated using acute lymphocytic leukaemia, both sexes, 9 SEER registries, 1990				Simulated using eye and orbit cancer and Vietnamese population, 9 SEER registries, 1990			
		Gamma method		Delta method		Gamma method		Delta method		Gamma method		Delta method	
		E_L	E_U	E_L	E_U	E_L	E_U	E_L	E_U	E_L	E_U	E_L	E_U
0	30	2.21	2.40	2.00	2.95	2.39	2.25	1.90	3.06	0.70	0	0	0
0	50	2.44	2.77	2.37	2.89	2.37	2.12	2.00	2.83	1.62	0	0	0
0	70	2.49	2.62	2.47	2.64	2.16	2.65	1.88	3.51	0.76	0	0	0
0	∞	2.64	2.43	2.63	2.49	2.26	2.04	1.90	3.24	0.75	0	0	0
30	50	2.38	2.28	2.31	2.36	2.04	2.04	1.39	3.99	0.32	0	0	0
30	70	2.48	2.37	2.45	2.43	1.91	2.16	1.38	3.72	0.70	0	0	0
30	∞	2.64	2.41	2.61	2.51	2.23	1.84	1.55	3.41	0.78	0	0	0
50	70	2.20	2.53	2.18	2.58	1.75	1.96	1.13	4.28	0.34	0	0	0
50	∞	2.30	2.33	2.28	2.40	1.95	1.63	1.45	3.62	0.76	0	0	0
70	∞	2.38	2.21	2.38	2.30	1.88	2.00	1.11	4.31	0.16	0	0	0

E_L = per cent of $L >$ TRUE and E_U = per cent of $U <$ TRUE.

x. We assumed that $\lambda_c(a)$ is constant within an interval and computed $\lambda_c^*(a)$ which is not constant. However, it may have been more realistic to assume the hazards among the actual at risk populations, $\lambda_c^*(a)$, are constant over the interval and computed the non-constant $\lambda_c(a)$. Unfortunately, this approach does not appear to be tractable.

We have generalized the gamma confidence intervals [1] to apply to our new statistic. Although these intervals appear conservative in cases with extremely low counts, we have shown that the delta method which adds 0.5 to zero counts in the estimation of the variance of the counts performs worse. A more general way of performing the delta method is to assume that variances associated with zero counts are equal to some constant, $0<\delta<1$. The problem is that there is no obvious choice of δ; we have arbitrarily chosen $\delta=0.5$ in this paper. Note in the most extreme case where all counts are zero, the generalized gamma interval gives a non-zero upper limit, while the delta method gives an upper limit that approaches zero as $\delta\rightarrow 0$. Other methods, such as parametric bootstrap confidence intervals, suffer from the same problem of having no satisfactory method for handling zero counts. In the simple case of linear combinations of independent Poisson variables, Fay and Feuer [1] discuss similar issues comparing the gamma intervals and the approximate bootstrap confidence (ABC) intervals.

Our new estimator happens to be numerically similar to the existing method of Wun *et al.* [2]. Because our approach is new and not simply a modification of Wun *et al.* [2] and because the notations are very different between the two approaches, we have relegated the full comparison between the two methods to a technical report [7]. In that report, we show through Taylor series approximations that the two methods are similar. In addition, using the new method we recalculated Table I-17 of the *Cancer Statistics Review*, *1973–1998* [6] which gives lifetime risk of developing cancer calculated for each of 30 difference cancer categories on six subpopulations, and the new estimator differs by less than 2 per cent from that of Wun *et al.* [2] in every case (see [7]).

DEVCAN (Probability of DEVeloping CANcer) software [8] has been freely available to calculate the statistics of Wun *et al.* [2] and will be updated to calculate our new estimator in a future version. See http://srab.cancer.gov/DevCan/ for the most current version of the software.

APPENDIX A: WRITING $A(x,y,z)$ AS A SUM

We write our estimator of $A(x,y)$ as $A(x,y,z)$ (see Section 3.2 for motivation of this notation).

Let $a_i\leqslant x<a_{i+1}$ and $a_j<y\leqslant a_{j+1}$ for $x<y, i\leqslant j$, and $j\leqslant k$. For convenience we regroup the ages after inserting group delimiters at x and y. Let the new delimiters be $0=b_0\leqslant b_1\leqslant b_2\leqslant\cdots\leqslant b_{k+3}=\infty$ where $b_0=a_0,\dots,b_i=a_i, b_{i+1}=x, b_{i+2}=a_{i+1},\dots,b_{j+1}=a_j, b_{j+2}=y, b_{j+3}=a_{j+1},\dots,$ $b_{k+3}=a_{k+1}=\infty$. We let

$$\hat{S}(b_\ell)=\exp\left\{-\int_0^{b_\ell}\hat{\lambda}(u)\,\mathrm{d}u\right\}=\exp\left\{-\sum_{u=0}^{\ell-1}\hat{\lambda}(b_u)(b_{u+1}-b_u)\right\}$$

and similarly $\hat{S}_{\rm d}(b_\ell) = \exp\{-\int_0^{b_\ell} \hat{\lambda}_{\rm d}(u)\,{\rm d}u\}$ and $\hat{S}^*_{\rm o}(b_\ell) = \exp\{-\int_0^{b_\ell} \hat{\lambda}^*_{\rm o}(u)\,{\rm d}u\}$. In this notation, $A(x, y) = A(b_{i+1}, b_{j+2})$, and we estimate it with

$$A(b_{i+1}, b_{j+2}, \boldsymbol{z}) = \frac{\sum_{\ell=i+1}^{j+1} \int_{b_\ell}^{b_{\ell+1}} \hat{\lambda}_{\rm c}(b_\ell)\hat{S}(b_\ell)\, \exp(-\int_{b_\ell}^{u} \hat{\lambda}(b_\ell)\,{\rm d}t)\,{\rm d}u}{\hat{S}_{\rm o}(b_{i+1})\{1 - \sum_{\ell=0}^{i} \int_{b_\ell}^{b_{\ell+1}} \hat{\lambda}_{\rm c}(b_\ell)\hat{S}_{\rm d}(b_\ell)\, \exp(-\int_{b_\ell}^{u} \hat{\lambda}_{\rm d}(b_\ell)\,{\rm d}t)\,{\rm d}u\}}$$

$$= \frac{\sum_{\ell=i+1}^{j+1} \hat{\lambda}_{\rm c}(b_\ell)\hat{S}(b_\ell) \int_{b_\ell}^{b_{\ell+1}} \exp(-(u - b_\ell)\hat{\lambda}(b_\ell))\,{\rm d}u}{\hat{S}_{\rm o}(b_{i+1})\{1 - \sum_{\ell=0}^{i} \hat{\lambda}_{\rm c}(b_\ell)\hat{S}_{\rm d}(b_\ell) \int_{b_\ell}^{b_{\ell+1}} \exp(-(u - b_\ell)\hat{\lambda}_{\rm d}(b_\ell))\,{\rm d}u\}}$$

Because $\hat{\lambda}(b_\ell)$ or $\hat{\lambda}_{\rm d}(b_\ell)$ may equal zero and $b_{\ell+1}$ may equal infinity, we let $\phi(\lambda, \ell) = \int_{b_\ell}^{b_{\ell+1}} \exp(-(u - b_\ell)\lambda)\,{\rm d}u$. These integrals are

$$\phi(\lambda, \ell) = \begin{cases} \frac{1-\exp[-(b_{\ell+1}-b_\ell)\lambda]}{\lambda} & \text{if } \lambda > 0 \text{ and } b_{\ell+1} \neq \infty \\ b_{\ell+1} - b_\ell & \text{if } \lambda = 0 \text{ and } b_{\ell+1} \neq \infty \\ \frac{1}{\lambda} & \text{if } \lambda > 0 \text{ and } b_{\ell+1} = \infty \\ \infty & \text{if } \lambda = 0 \text{ and } b_{\ell+1} = \infty \end{cases}$$

where the case $\lambda = 0$ and $b_{\ell+1} = \infty$ is one of the 'impossible' hypothetical cohorts (see Section 3.1). Thus, we obtain

$$A(b_{i+1}, b_{j+2}, \boldsymbol{z}) = \frac{\sum_{\ell=i+1}^{j+1} \hat{\lambda}_{\rm c}(b_\ell)\hat{S}(b_\ell)\phi(\hat{\lambda}(b_\ell), \ell)}{\hat{S}^*_{\rm o}(b_{i+1})\{1 - \sum_{\ell=0}^{i} \hat{\lambda}_{\rm c}(b_\ell)\hat{S}_{\rm d}(b_\ell)\phi(\hat{\lambda}_{\rm d}(b_\ell), \ell)\}}$$

APPENDIX B: ASYMPTOTIC BEHAVIOUR OF THE GAMMA INTERVALS

Fay and Feuer [1] stated that the gamma intervals approach the standard normal intervals if (using the application of this paper) $A(\boldsymbol{z})$ goes to infinity in such a way that $V(\boldsymbol{z})A(\boldsymbol{z})^{-1}$ remains constant. This is not helpful for our situation (nor is it particularly helpful for studying directly standardized rates as in Fay and Feuer [1]). Here assume that the mean counts, $\boldsymbol{\mu}$, and the person-years, $\boldsymbol{n} = \{n_0^{(\rm c)}, n_1^{(\rm c)}, \ldots, n_k^{(\rm o)}\}$, both increase by the same factor, say N. Since $A(\boldsymbol{\mu})$ is a function of the rates only, this value does not change as N increases; however, as one would expect, the variance estimates will change by a factor of N^{-1}. We write the lower confidence limit in terms of the chi-square distribution $\{V/(2NA)\}(\chi^2)^{-1}_{2NA^2/V}(\alpha/2)$, where $A = A(\boldsymbol{z})$ and $N^{-1}V = V(\boldsymbol{z})$. The difference of the lower gamma confidence limit and the standard normal lower limit approaches zero as $N \to \infty$:

$$\lim_{N\to\infty} \left\{ \frac{V}{2NA}(\chi^2)^{-1}_{\frac{2NA^2}{V}}(\alpha/2) - \left(A + \left(\frac{V}{N}\right)^{1/2} \Phi^{-1}(\alpha/2)\right)\right\}$$

$$= \lim_{N\to\infty} \left(\frac{V}{N}\right)^{1/2} \left\{ \frac{(\chi^2)^{-1}_{\frac{2NA^2}{V}}(\alpha/2) - \frac{2NA^2}{V}}{\sqrt{2}(\frac{2NA^2}{V})^{1/2}} - \Phi^{-1}(\alpha/2) \right\} = 0$$

where the result follows since $\lim_{v\to\infty} \frac{(\chi^2)^{-1}_v(p)-v}{\sqrt{2v}} = \Phi^{-1}(p)$ (Johnson and Kotz, reference [9, p. 170]), and $\Phi^{-1}(p)$ is the pth quantile of the standard normal distribution. One can similarly show that the upper confidence limits approach the standard normal limits.

ACKNOWLEDGEMENTS

We thank Chris Gunther and Alexander Korobkov for computing help.

REFERENCES

1. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. *Statistics in Medicine* 1997; **16**:791–801.
2. Wun L-M, Merrill RM, Feuer EJ. Estimating lifetime and age-conditional probabilities of developing cancer. *Lifetime Data Analysis* 1998; **4**:169–186.
3. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980; 163–178.
4. Brillinger DR. The natural variability of vital rates and associated statistics (with Discussion). *Biometrics* 1986; **42**:693–734.
5. Lehmann EL. *Elements of Large-Sample Theory*. Springer: New York, 1999.
6. Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, Edwards BK (eds). *SEER Cancer Statistics Review, 1973–1998*. National Cancer Institute: Bethesda, MD, 2001 (accessed at http://seer.cancer.gov/csr/1973_1998/ on 3 September 2002).
7. Fay MP, Pfeiffer R, Cronin KA, Le C, Feuer EJ. Comparison of two methods for calculating age-conditional probabilities of developing cancer. Technical report #2002–01, Statistical Research and Applications Branch, National Cancer Institute, 2002 (accessed at http://srab.cancer.gov/reports on 3 September 2002).
8. National Cancer Institute and Information Management Services. *DEVCAN*: *Probability of DEVeloping CANcer software* Version 4.1. National Cancer Institute and Information Management Services, Inc., 2001 (accessed at http://srab.cancer.gov/DevCan/ on 3 September 2002).
9. Johnson NL, Kotz S. *Distributions in Statistics*: *Continuous Univariate Distributions-1*. Wiley: New York, 1970.