Estimating Age-Conditional Probability of Developing Cancer using a Piecewise Mid-Age Joinpoint Model for the Rates. [1]

Michael P. Fay
National Cancer Institute
6116 Executive Blvd., Suite 504
Bethesda, MD 20892-8317
U.S.A.
faym@mail.nih.gov

April 14, 2004

[1]This report corrects and updates:

> Fay MP. Estimating age-conditional probability of developing cancer using a piecewise mid-age joinpoint model for the rates. Statistical Research and Applications Branch, NCI, Technical Report # 2003-03.

When citing this report use:

> Fay MP. Estimating age-conditional probability of developing cancer using a piecewise mid-age joinpoint model for the rates. Statistical Research and Applications Branch, NCI, Technical Report # 2003-03-A (http://srab.cancer.gov/reports).

Abstract

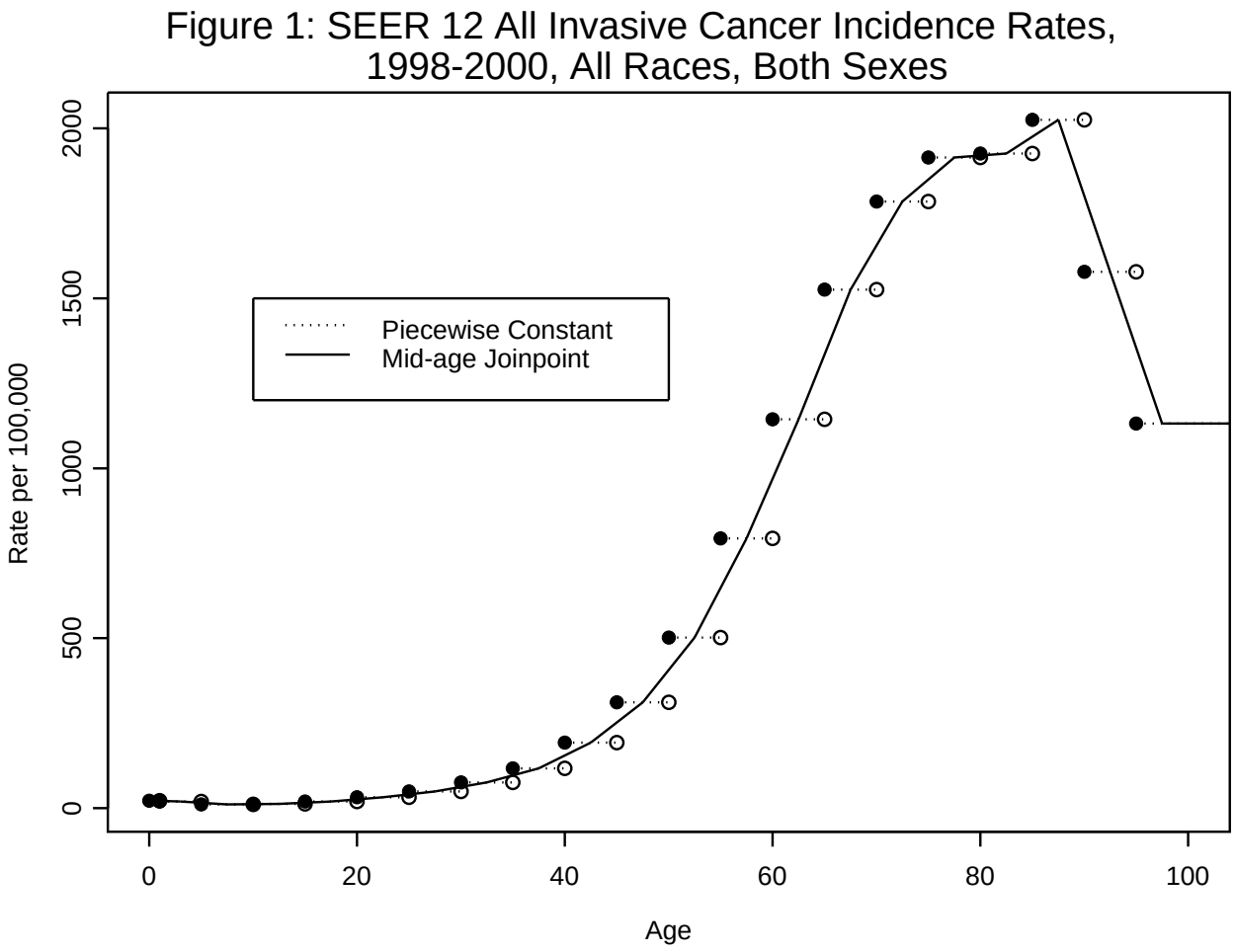
Fay, Pfeiffer, Cronin, Le, and Feuer (*Statistics in Medicine* 2003; **22;** 1837-1848) developed a formula to calculate the age-conditional probability of developing a disease for the first time (ACPDvD) for a hypothetical cohort. The novelty of the formula of Fay et al (2003) is that one need not know the rates of first incidence of disease per person-years alive *and disease free*, but may input the rates of first incidence per person-years alive only. The latter rates are much easier to estimate. Other inputs into the formula are the rates of death from the disease and rate of death from other causes per person-years alive. Fay et al. (2003) used simple piecewise constant models for all three rate functions which have constant rates within each age group. In this paper, we detail a method for estimating rate functions which does not have jumps at the beginning of age groupings, and need not be constant within age groupings. We call this method the mid-age group joinpoint (MAJ) model for the rates. The drawback of the MAJ model is that numerical integration must be used to estimate the resulting ACPDvD. To increase computational speed, we offer a piecewise approximation to the MAJ model, which we call the piecewise mid-age group joinpoint (PMAJ) model. The PMAJ model for the rates input into the formula for ACPDvD described in Fay et al. (2003) is the current method used in the freely available DevCan software made available by the National Cancer Institute.

1 Introduction

Fay, Pfeiffer, Cronin, Le, and Feuer (2003) showed how to calculate the age-conditional probabilities of developing a disease (ACPDvD) from registry data. Throughout this paper we use "cancer" as our disease of interest, but the method applies to specific types

of cancer as well as other diseases where information is collected by population based surveillance methods. Fay et al. (2003) provided a formula (see equation 1 below) to calculate ACPDvD after inputing the rate function by age of (1) first incidence of cancer per person-years alive, (2) death from cancer per person-years alive, and (3) death from other causes per person-years alive. Fay et al. (2003) used a simple piecewise constant model for the three rate functions, which have constant rates within each age group. Here we detail two more complicated models for the rates. The first model is a segmented regression model or joinpoint model for the rates, where the rate function is a series of linear functions that join at the mid-points of the age groups, and the rate function is constant before the first mid-point and after the last "mid-point" (because the last interval goes to infinity, the last "mid-point" is not really a mid-point at all, see below). We will call this model the MAJ (mid-age group joinpoint) model for the rates. In Figure 1 we show how both the piecewise constant model and the mid-age group joinpoint model apply to all invasive cancer incidence from the Surveillance Epidemiology and End Results (SEER) program of the U.S. National Cancer Institute in 1998-2000. Figure 1 uses the SEER 12 registries which cover about 14 percent of the U.S. population, covering 5 states (Connecticut, Hawaii, Iowa, New Mexico, Utah), 6 metropolitan areas (Atlanta, Detroit, Los Angeles, San Francisco-Oakland, San Jose-Monterey, Seattle-Puget Sound) and the Alaska Native Registry (see Ries, et al. 2003).

Notice that the MAJ model gives a more smoothly changing and probably a better modeled rate. The only place where the MAJ model may not perform better than the piecewise constant model is at peaks or valleys, where there may be some bias. In Figure 1 we see that the smoothness of the MAJ appears to produce more plausible estimates for ages 0 through 85 and from ages 90 and above, and the only age group with a noteworthy bias problem is 85 to 90. Thus, for almost all of the age range the



MAJ model is more plausible.

A problem with the mid-age group joinpoint model is that it requires numeric integration for its calculation. A faster method uses a series of piecewise constant values to approximate the mid-age group joinpoint model. We call this the PMAJ (piecewise mid-age group joinpoint) model. The PMAJ does not require numeric integration, so it is much faster than the MAJ model. The PMAJ model is a piecewise constant model that differs from the piecewise constant model of Fay et al (2003) in that the pieces are smaller and the corresponding values of the rates are motivated by the MAJ model.

Starting with version 5.0, the freely available DevCan software (DevCan, 2003) uses the PMAJ method. (There was a small calculation error in versions 5.0 and 5.1 that will be corrected by version 5.2). DevCan calculates ACPDvD or age conditional probability of dying from a disease for U.S. cancer data or for user supplied data.

The outline of this paper is as follows. Section 2 gives the motivation for the MAJ estimator of age-conditional probability of developing cancer. Appendix A shows how to calculate the integral needed for Section 2. Section 3 describes the PMAJ model and how it is used to estimate the age-conditional probability of developing cancer. Section 4 gives an example of the estimator of ACPDvD using three different methods for estimating the rates, the simple piecewise constant method proposed in Fay et al. (2003), the MAJ method, and the PMAJ method. For completeness a second Appendix, Appendix B, compares the PMAJ method with the method of Wun, et al. (1998), since the latter method was the method used by previous versions of the DevCan software.

2 Mid-Age group Joinpoint Estimator

Fay, et al. (2003) assumed that the hazard rate for other cause (i.e., non-cancer) mortality is the same for people with and without cancer. Fay et al. (2003) gave a formula for the age-conditional probability of developing cancer between the ages of x and y given alive and cancer-free just before age x as

$$A(x,y) \;=\; \frac{\int_x^y \lambda_c(u) S_a(u-) du}{S_o(x-)\left\{1 - \int_0^x \lambda_c(u) S_d(u-) du\right\}}. \qquad (1)$$

See Table 1 for the notation taken from Fay, et al. (2003). The only change in notation from Fay, et al. (2003) is that we use the subscript a to represent all causes of events instead of a blank subscript. For example, we let $S^*(u) = S^*_a(u)$. Other notation in this paper is defined as it is introduced.

Table 1: Notation

Random Variables and Parameters	
$T=$ age at death	$T^*=$ age at first cancer or death before cancer
$J=$ type of death $(J=d)$=death from cancer $(J=o)$=death from other causes	$J^*=$ type of event $(J^*=c)$=first cancer $(J^*=o)$=death before first cancer
$\lambda_c(t)=$ rate at t for first cancer given alive	$\lambda_c^*(t)=$ rate at t for first cancer given alive and cancer-free
$\lambda_o(t)=$ rate at t for death before cancer given alive	$\lambda_o^*(t)=$ rate at t for death before cancer given alive and cancer-free
$\lambda_d(t)=$ rate at t for death from cancer given alive	
$\lambda_a(t)=$ rate at t for death given alive	$\lambda_a^*(t)=$ rate at t for first cancer or death before first cancer given alive and cancer-free
$S_j(t)=\exp\left\{-\int_0^t \lambda_j(u)\,du\right\}$ for $j=a,c,o,d$	$S_j^*(t)=\exp\left\{-\int_0^t \lambda_j^*(u)\,du\right\}$ for $j=a,c,o$
Observations	
Within the age interval, $[a_i, a_{i+1})$, and within the calendar interval of interest we observe...	
$c_i=$ number of first cancer incident cases	$n_i^{(j)}=$ estimate of person-years alive associated with $j=c,d,o$ (DevCan uses the sum of mid-year populations during the calendar interval of interest)
$d_i=$ number of cancer deaths	
$o_i=$ number of other deaths	

In Fay et al (2003), the rates were estimated by a piecewise constant model. Here we use a mid-age group joinpoint (MAJ) model, where we draw lines connecting the midpoints of the intervals except the first and last interval. The first interval is constant until the midpoint, and the last interval is constant after a nominal "midpoint". This nominal "midpoint" is half the length of the previous age interval from the beginning of the last interval, and would be the midpoint if the last age interval was the same length as the previous interval.

We introduce new notation for breaking up the ages. Fay, et al. (2003) used $0 = a_0 < a_1 < \cdots < a_k < a_{k+1} = \infty$. Here we use a joinpoint model with joins at the midpoints (and nominal midpoint),

$$\frac{a_1}{2} < \frac{a_1 + a_2}{2} < \cdots < \frac{a_{k-1} + a_k}{2} < a_k + \frac{a_k - a_{k-1}}{2}.$$

Let

$$0 = t_{-1} < t_0 = \frac{a_1}{2} < t_1 = \frac{a_1 + a_2}{2} < \cdots < t_{k-1} = \frac{a_{k-1} + a_k}{2} < t_k = a_k + \frac{a_k - a_{k-1}}{2} < t_{k+1} = \infty$$

(The indices start at -1 so that the index values for the rate estimators, $\tilde{\lambda}_{ji}$, match up with the count notation of Fay et al., 2003.) The MAJ estimator for the rate of event j (for $j = c, d$, or o) at t_i (for $i = 0, 1, \ldots, k$) is

$$\tilde{\lambda}_{ji} = \tilde{\lambda}_j(t_i) = \frac{j_i}{n_i^{(j)}}, \tag{2}$$

where j_i is either c_i, d_i, or o_i as defined in Table 1. (Note that $\tilde{\lambda}_j(t_i) = \hat{\lambda}_j(a_i) = \hat{\lambda}_j(t_i)$, where $\hat{\lambda}_j(\cdot)$ is the piecewise constant function used by Fay et al. [2003]). We define $\tilde{\lambda}_{j,-1} = \tilde{\lambda}_{j0}$ and $\tilde{\lambda}_{j,k+1} = \tilde{\lambda}_{jk}$. For $j = a$, MAJ estimator for the rate at t_i is

$$\tilde{\lambda}_{ai} = \tilde{\lambda}_a(t_i) = \frac{o_i}{n_i^{(o)}} + \frac{d_i}{n_i^{(d)}}. \tag{3}$$

Then for $t \in [t_i, t_{i+1})$ for $i = 1, \ldots, k$, we define $\tilde{\lambda}_j(t)$ as the point on the line defined by connecting the points $(t_i, \tilde{\lambda}_{ji})$ and $(t_{i+1}, \tilde{\lambda}_{j,i+1})$. In other words,

$$\tilde{\lambda}_j(t) = \alpha_{ji} + \beta_{ji} t,$$

where

$$\alpha_{ji} = \frac{t_{i+1}\tilde{\lambda}_{ji} - t_i \tilde{\lambda}_{j,i+1}}{t_{i+1} - t_i} \tag{4}$$

(5)

and

$$\beta_{ji} = \left(\frac{\tilde{\lambda}_{j,i+1} - \tilde{\lambda}_{j,i}}{t_{i+1} - t_i} \right). \tag{6}$$

Thus, $\alpha_{j,-1} = \tilde{\lambda}_{j0}$ and $\beta_{j,-1} = 0$, and similarly by taking limits as $t_{k+1} \to \infty$ then $\alpha_{j,k} = \tilde{\lambda}_{j,k}$ and $\beta_{j,k} = 0$.

Now $\tilde{S}_j(u)$ for $u \in [t_i, t_{i+1})$ is

$$\begin{aligned} \tilde{S}_j(u) &= \exp\left(- \int_0^u \tilde{\lambda}_j(t) dt \right) \\ &= \exp\left(- \sum_{\ell=0}^{i} \int_{t_{\ell-1}}^{t_\ell} \{\alpha_{j,\ell-1} + \beta_{j,\ell-1} t\} \, dt - \int_{t_i}^{u} \{\alpha_{j,i} + \beta_{j,i} t\} \, dt \right) \end{aligned}$$

Note that (for $\ell = 0, 1, \ldots, k$)

$$\begin{aligned} \int_{t_{\ell-1}}^{t_\ell} \{\alpha_{j,\ell-1} + \beta_{j,\ell-1} t\} \, dt &= (t_\ell - t_{\ell-1}) \alpha_{j,\ell-1} + (t_\ell^2 - t_{\ell-1}^2) \frac{\beta_{j,\ell-1}}{2} \\ &= t_\ell \tilde{\lambda}_{j,\ell-1} - t_{\ell-1} \tilde{\lambda}_{j,\ell} + (t_\ell - t_{\ell-1})(t_\ell + t_{\ell-1}) \frac{\beta_{j,\ell-1}}{2} \\ &= t_\ell \tilde{\lambda}_{j,\ell-1} - t_{\ell-1} \tilde{\lambda}_{j,\ell} + (t_\ell + t_{\ell-1}) \left(\frac{\tilde{\lambda}_{j,\ell} - \tilde{\lambda}_{j,\ell-1}}{2} \right) \\ &= (t_\ell - t_{\ell-1}) \left(\frac{\tilde{\lambda}_{j,\ell-1} + \tilde{\lambda}_{j,\ell}}{2} \right) \end{aligned}$$

so that for $i = 0, 1, \ldots, k$,

$$\tilde{S}_j(t_i) \;=\; \exp\left(-\sum_{\ell=0}^{i}(t_\ell - t_{\ell-1})\left(\frac{\tilde{\lambda}_{j,\ell-1}+\tilde{\lambda}_{j,\ell}}{2}\right)\right)$$

Also notice that (when $u < \infty$)

$$\int_{t_i}^{u}\{\alpha_{j,i}+\beta_{j,i}t\}\,dt \;=\; (u-t_i)\alpha_{j,i} + (u^2-t_i^2)\frac{\beta_{j,i}}{2}$$

Therefore when $u \in [t_i, t_{i+1})$,

$$\begin{aligned}\tilde{S}_j(u) &= \exp\left(-\sum_{\ell=0}^{i}(t_\ell - t_{\ell-1})\left(\frac{\tilde{\lambda}_{j,\ell-1}+\tilde{\lambda}_{j,\ell}}{2}\right) - (u-t_i)\alpha_{j,i} - (u^2-t_i^2)\frac{\beta_{j,i}}{2}\right)\\ &= \tilde{S}_j(t_i)\exp\left(-\left[(u-t_i)\alpha_{j,i} + (u^2-t_i^2)\frac{\beta_{j,i}}{2}\right]\right)\end{aligned}$$

Let $\tilde{A}(x,y)$ be the estimator of $A(x,y)$ using the MAJ model. The two integrals we need to estimate for $\tilde{A}(x,y)$ are of the type,

$$\tilde{F}_{j,h}(t) = \int_0^t \tilde{\lambda}_j(u)\tilde{S}_h(u-)du, \tag{7}$$

where in the numerator of $\tilde{A}(x,y)$ we need $\tilde{F}_{c,a}$ (i.e., $j = c$ and $h = a$ in equation 7), and in the denominator of $\tilde{A}(x,y)$ we need $\tilde{F}_{c,d}$. Suppose, without loss of generality, that $t \in [t_i, t_{i+1})$, then

$$\begin{aligned}\tilde{F}_{j,h}(t) &= \sum_{\ell=-1}^{i-1}\int_{t_\ell}^{t_{\ell+1}}\tilde{\lambda}_j(u)\tilde{S}_h(u-)du + \int_{t_i}^{t}\tilde{\lambda}_j(u)\tilde{S}_h(u-)du\\ &= \sum_{\ell=-1}^{i-1}\tilde{S}_h(t_\ell)\int_{t_\ell}^{t_{\ell+1}}(\alpha_{j\ell}+\beta_{j\ell}u)\exp\left(-\left[(u-t_\ell)\alpha_{h\ell} + (u^2-t_\ell^2)\frac{\beta_{h\ell}}{2}\right]\right)du\\ &\quad +\tilde{S}_h(t_i)\int_{t_i}^{t}(\alpha_{ji}+\beta_{ji}u)\exp\left(-\left[(u-t_\ell)\alpha_{hi} + (u^2-t_\ell^2)\frac{\beta_{hi}}{2}\right]\right)du\\ &= \sum_{\ell=-1}^{i-1}\tilde{S}_h(t_\ell)R_{j,h}(t_\ell, t_{\ell+1}) + \tilde{S}_h(t_i)R_{j,h}(t_i, t)\end{aligned}$$

where $R_{j,h}(t_\ell, v)$ (for $\ell = -1, 0, 1, 2, \ldots, i$ and $v \leq t_{\ell+1}$) is defined implicitly (see the Appendix). Then,

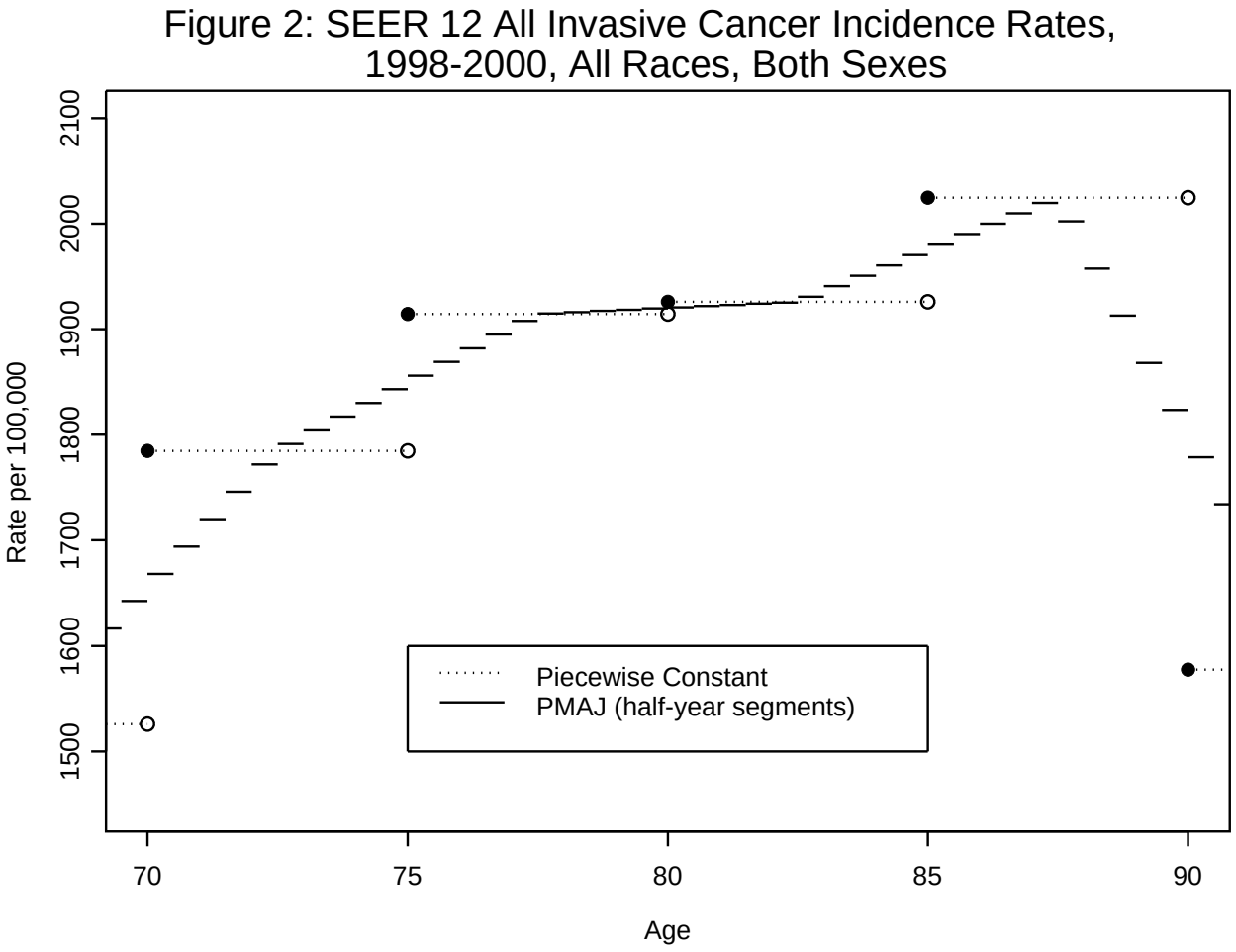
$$\tilde{A}(x,y) \;=\; \frac{\tilde{F}_{c,a}(y) - \tilde{F}_{c,a}(x)}{\tilde{S}_o(x)\left\{1 - \tilde{F}_{c,d}(x)\right\}}.$$

3 Piecewise Mid-Age group Joinpoint Estimator

In the MAJ model we divided up the age line into $k+2$ intervals. Here we define those intervals in both the t_i notation and the a_i notation.

$$\begin{aligned}
I_0 &= [t_{-1}, t_0) = \left[0, \frac{a_1}{2}\right) \\
I_1 &= [t_0, t_1) = \left[\frac{a_1}{2}, \frac{a_1 + a_2}{2}\right) \\
&\vdots \qquad \vdots \qquad \vdots \\
I_i &= [t_{i-1}, t_i) = \left[\frac{a_{i-1} + a_i}{2}, \frac{a_i + a_{i+1}}{2}\right) \\
&\vdots \qquad \vdots \qquad \vdots \\
I_k &= [t_{k-1}, t_k) = \left[\frac{a_{k-1} + a_k}{2}, a_k + \frac{a_k - a_{k-1}}{2}\right) \\
I_{k+1} &= [t_k, \infty) = \left[a_k + \frac{a_k - a_{k-1}}{2}, \infty\right)
\end{aligned}$$

In the MAJ model the rates for the first and the last intervals are represented by lines with zero slope, and the rates for the ith interval ($i = 1, \ldots, k$) for the jth rate type ($j = a, c, d, o$) is a line defined by connecting the points $(t_{i-1}, \tilde{\lambda}_{j,i-1})$ and $(t_i, \tilde{\lambda}_{ji})$ (see equations 2 and 3 for definition of $\tilde{\lambda}_{ji}$). In the PMAJ model we divide the ith interval into m_i equal sized intervals, and use a piecewise constant estimate on each of those m_i intervals. One way to define m_i is to chose m_i so that each equal sized interval is 1/2 year long. In other words, $m_i = 2(t_i - t_{i-1})$. This is the definition of m_i that we use for the DevCan software (starting with version 5.0, see DevCan, 2003), but all the



following holds for arbitrary m_i. In Figure 2 we show the PMAJ model with half-year intervals and the piecewise constant model for the US all invasive cancer mortality rates for ages 70 through 90 years.

Here are the details. Consider the hth (for $h = 1, \ldots, m_i$) of the m_i intervals within interval i (for $i = 1, \ldots, k$) for rate type j (for $j = a, c, d, o$). This interval is

$$\left[t_{i-1} + \frac{(h-1)(t_i - t_{i-1})}{m_i}, t_{i-1} + \frac{h \cdot (t_i - t_{i-1})}{m_i} \right)$$

For convenience we introduce new notation for the ends of this interval, let

$$t_{i-1,h} = t_{i-1} + \frac{h \cdot (t_i - t_{i-1})}{m_i}$$

so that $t_{i-1,0} = t_{i-1}$ and $t_{i-1,m_i} = t_i$. At the beginning of this interval the value of the rate is

$$\begin{aligned}
\tilde{\lambda}_j\left(t_{i-1,h-1}\right) &= \alpha_{j,i-1} + \beta_{j,i-1}\left(t_{i-1} + \frac{(h-1)(t_i - t_{i-1})}{m_i}\right) \\
&= \frac{t_i\tilde{\lambda}_{j,i-1} - t_{i-1}\tilde{\lambda}_{ji}}{t_i - t_{i-1}} + \frac{(\tilde{\lambda}_{ji} - \tilde{\lambda}_{j,i-1})t_{i-1}}{t_i - t_{i-1}} + \frac{(h-1)(\tilde{\lambda}_{ji} - \tilde{\lambda}_{j,i-1})}{m_i} \\
&= \tilde{\lambda}_{j,i-1} + \frac{(h-1)(\tilde{\lambda}_{ji} - \tilde{\lambda}_{j,i-1})}{m_i}
\end{aligned}$$

(see equations 4 and 6 for definitions of $\alpha_{j,i-1}$ and $\beta_{j,i-1}$). Similarly at the end of this interval the rate is

$$\tilde{\lambda}_j\left(t_{i-1,h}\right) = \tilde{\lambda}_{j,i-1} + \frac{h(\tilde{\lambda}_{ji} - \tilde{\lambda}_{j,i-1})}{m_i}$$

For the PMAJ model we simply assume a constant rate equal to the average of the beginning and the end values of the rate over this interval. In other words, under the PMAJ model for any $t \in [t_{i-1,h-1}, t_{i-1,h})$ we estimate the rate with

$$\dot{\lambda}_j(t) = \tilde{\lambda}_{j,i-1} + \frac{(2h-1)(\tilde{\lambda}_{ji} - \tilde{\lambda}_{j,i-1})}{2m_i}$$

Since the PMAJ model is a piecewise model, we can use Appendix A of Fay *et al.* (2003) to express the estimator of age conditional probability of developing cancer. The only hard part is correctly defining the starting and ending of each piecewise interval. The ends of these intervals are

$$0 \equiv t_{-1} < t_0 < t_{0,1} < t_{0,2} < \cdots < t_{0,m_1-1} < t_1 < t_{1,1} < \cdots < t_{k-1,m_k-1} < t_k < t_{k+1} \equiv \infty$$

For convenience write these interval ends with only a single index as

$$0 \equiv \tau_0 < \tau_1 < \tau_2 < \tau_3 < \cdots < \tau_{m_1} < \tau_{m_1+1} < \tau_{m_1+2} < \cdots < \tau_{M-1} < \tau_M < \tau_{M+1} \equiv \infty$$

where $M = \sum_{i=0}^{k} m_i$, and $m_0 = 1$. In other words, $t_{-1} = \tau_0$ and for $i = 0, 1, \ldots, k$, then $t_i = \tau_{g(i)}$ and $t_{i,h} = \tau_{g(i)+h}$, where $g(i) = \sum_{\ell=0}^{i} m_\ell$.

Now we can follow very similar notation to Appendix A of Fay *et al.* (2003). We now repeat that Appendix with the modifications to notation required for the PMAJ model. Let the estimator of $A(x, y)$ under the PMAJ model be denoted $\dot{A}(x, y)$. Let $\tau_i \le x < \tau_{i+1}$ and $\tau_j < y \le \tau_{j+1}$ for $x < y, i \le j$, and $j \le M + 2$. For convenience we regroup the ages after inserting group delimiters at x and y. Let the new delimiters be $0 = b_0 \le b_1 \le b_2 \le \cdots \le b_{M+3} = \infty$ where $b_0 = \tau_0, \ldots, b_i = \tau_i, b_{i+1} = x, b_{i+2} = \tau_{i+1}, \ldots, b_{j+1} = \tau_j, b_{j+2} = y, b_{j+3} = \tau_{j+1}, \ldots, b_{M+3} = \tau_{M+1} = \infty$. We let

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

and similarly $\dot{S}_d(b_\ell) = \exp\left\{-\int_0^{b_\ell} \dot{\lambda}_d(u)du\right\}$ and $\dot{S}_o(b_\ell) = \exp\left\{-\int_0^{b_\ell} \dot{\lambda}_o(u)du\right\}$. In this notation, the probability of developing cancer by age y given survival until age x is $A(x, y) = A(b_{i+1}, b_{j+2})$, and under the PMAJ model we estimate it with

$$\begin{aligned}
\dot{A}(b_{i+1}, b_{j+2}) &= \frac{\sum_{\ell=i+1}^{j+1} \int_{b_\ell}^{b_{\ell+1}} \dot{\lambda}_c(b_\ell)\dot{S}_a(b_\ell)\exp\left(-\int_{b_\ell}^{u} \dot{\lambda}_a(b_\ell)dt\right)du}{\dot{S}_o(b_{i+1})\left\{1 - \sum_{\ell=0}^{i} \int_{b_\ell}^{b_{\ell+1}} \dot{\lambda}_c(b_\ell)\dot{S}_d(b_\ell)\exp\left(-\int_{b_\ell}^{u} \dot{\lambda}_d(b_\ell)dt\right)du\right\}} \\
&= \frac{\sum_{\ell=i+1}^{j+1} \dot{\lambda}_c(b_\ell)\dot{S}_a(b_\ell)\int_{b_\ell}^{b_{\ell+1}} \exp\left(-(u - b_\ell)\dot{\lambda}_a(b_\ell)\right)du}{\dot{S}_o(b_{i+1})\left\{1 - \sum_{\ell=0}^{i} \dot{\lambda}_c(b_\ell)\dot{S}_d(b_\ell)\int_{b_\ell}^{b_{\ell+1}} \exp\left(-(u - b_\ell)\dot{\lambda}_d(b_\ell)\right)du\right\}}.
\end{aligned}$$

Because $\dot{\lambda}_a(b_\ell)$ or $\dot{\lambda}_d(b_\ell)$ may equal zero and $b_{\ell+1}$ may equal infinity, we let $\phi(\lambda, \ell) = \int_{b_\ell}^{b_{\ell+1}} \exp\left(-(u - b_\ell)\lambda\right)du$. These integrals are

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

where the case $\lambda = 0$ and $b_{\ell+1} = \infty$ is one of the "impossible" hypothetical cohorts (see Section 3.1 of Fay *et al.* 2003). Thus, we obtain,

$$\dot{A}(b_{i+1}, b_{j+2}) \quad = \quad \frac{\sum_{\ell=i+1}^{j+1} \dot{\lambda}_c(b_\ell)\dot{S}_a(b_\ell)\phi(\dot{\lambda}_a(b_\ell), \ell)}{\dot{S}_o(b_{i+1})\left\{1 - \sum_{\ell=0}^{i} \dot{\lambda}_c(b_\ell)\dot{S}_d(b_\ell)\phi(\dot{\lambda}_d(b_\ell), \ell)\right\}}.$$

4 Examples and Discussion

In this section we explore several different methods for estimating the rate functions, all using the formula of Fay et al. (2003) (e.g., all using equation 1). This comparison explores the differences between the piecewise constant method proposed in Fay et al. (2003), the PMAJ method, and the MAJ method. A different comparison emphasizing differences between versions of the DevCan software is described in Appendix B.

For all of the examples we use data from 1998-2000 (see reference for SEER DevCan database, 2003). The incidence data come from the Surveillance, Epidemiology, and End Results (SEER) program of the (U.S.) National Cancer Institute, and mortality data from the (U.S.) National Center for Health Statistics. We use the SEER 12 registries which cover about 14 percent of the U.S. population. We only use the mortality data covering the same area as the SEER 12 registries cover. Because the SEER 12 registries have complete coverage only back through 1992, we only look back in the database until 1992 to delete any incident case that had previously been diagnosed with the cancer of interest. These incident cases are deleted so that they are not counted when estimating the counts of first cancer incidence (the c_i values). The mid-year population estimates (the n_i values) come from the sum U.S. Census estimates of mid-year populations from 1998, 1999, and 2000 for the SEER 12 catchment areas for the appropriate sex group (e.g., males for prostate cancer).

In Table 2 we show the results for all invasive cancers and acute lymphocytic

Table 2: Age Conditional Probability of Developing Different Types of Invasive Cancers (in Percent) from SEER 12, 1998-2000

Start Age	End Age	Model	All Invasive (Both Sexes)	Prostate (Male)	Breast (Female)	Acute Lymphocytic Leukemia (Both Sexes)
0	20	Piecewise const	0.3158	0.0009	0.0015	0.0669
		PMAJ, interval=.5	0.3260	0.0011	0.0021	0.0633
		MAJ	0.3260	0.0011	0.0021	0.0633
0	50	Piecewise const	4.0690	0.2002	1.9188	0.0837
		PMAJ, interval=.5	4.1657	0.2550	1.9492	0.0808
		MAJ	4.1657	0.2550	1.9492	0.0808
40	50	Piecewise const	2.5260	0.2032	1.5131	0.0053
		PMAJ, interval=.5	2.5976	0.2579	1.5169	0.0055
		MAJ	2.5975	0.2579	1.5169	0.0055
0	Inf	Piecewise const	42.0876	17.4952	13.6471	0.1154
		PMAJ, interval=.5	41.7547	17.3375	13.5477	0.1121
		MAJ	41.7574	17.3389	13.5485	0.1121
60	61	Piecewise const	1.2340	0.5989	0.3822	0.0009
		PMAJ, interval=.5	1.0852	0.4946	0.3627	0.0009
		MAJ	1.0852	0.4946	0.3627	0.0009
64	65	Piecewise const	1.2758	0.6131	0.3872	0.0009
		PMAJ, interval=.5	1.4453	0.7440	0.4045	0.0010
		MAJ	1.4453	0.7440	0.4045	0.0010
60	65	Piecewise const	6.0331	2.9128	1.8777	0.0042
		PMAJ, interval=.5	6.0622	2.9492	1.8758	0.0044
		MAJ	6.0622	2.9492	1.8759	0.0044

leukemia for both sexes, prostate cancer for males, and breast cancer for females. We see the PMAJ values approximate the MAJ values very well.

In conclusion, we have described several methods for estimating rates for input into a formula to calculate ACPDvD, and we have shown that the PMAJ method provides a fast and reasonable estimators for the rates.

Acknowledgements

I would like to thank Kathy Cronin for suggesting the PMAJ method and thank her and Ram Tiwari for reading and commenting on drafts of this article.

References

DEVCAN: Probability of DEVeloping CANcer software Version 5.1, Statistical Research and Applications Branch, National Cancer Institute, 2003. (http://srab.cancer.gov/DevCan/).

Fay, M.P., Pfeiffer, R., Cronin, K.A., Le, C. and Feuer, E.J. 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. (http://srab.cancer.gov/reports).

Fay, M.P., Pfeiffer, R., Cronin, K.A., Le, C.,Feuer, E.J. (2003). Age-Conditional Probabilities of Developing Cancer. *Statistics in Medicine* **22**(11) 1837-1848.

Lange, K. (1999). *Numerical Analysis for Statisticians* Springer:New York.

Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, Mariotto A, Fay MP, Feuer EJ, Edwards BK (eds). SEER Cancer Statistics Review, 1975-2000, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2000, 2003.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) DevCan database: SEER 12 Incidence and Mortality, 1993-2000, Follow-back year=1992 National Cancer Institute, DCCPS, Surveillance Research Program,

Cancer Statistics Branch, released April 2003, based on the November 2002 submission. Underlying mortality data provided by NCHS (www.cdc.gov/nchs).

Wun, L-M, Merrill, R.M., and Feuer, E.J. Estimating lifetime and age-conditional probabilities of developing cancer. *Lifetime Data Analysis* 1998; **4,** 169-186.

A Calculation of R function

Recall that $R_{j,h}(t_\ell, v)$ represents an integral with 4 parameters. We can write it as

$$R(t_\ell, v, \alpha_{j\ell}, \beta_{j\ell}, \alpha_{h\ell}, \beta_{h\ell}) = \int_{t_\ell}^{v} (\alpha_{j\ell} + \beta_{j\ell} x) \exp\left(- \left[(x - t_\ell)\alpha_{h\ell} + (x^2 - t_\ell^2)\frac{\beta_{h\ell}}{2} \right] \right) dx$$

To simplify notation substitute let $t_\ell = u$ and $\alpha_{j\ell} = a_j$,$\beta_{j\ell} = b_j$,$\alpha_{h\ell} = a_h$, and $\beta_{h\ell} = b_h$. Thus,

$$R(u, v, a_j, b_j, a_h, b_h) = \int_{u}^{v} (a_j + b_j x) \exp\left(- \left[(x - u)a_h + (x^2 - u^2)\frac{b_h}{2} \right] \right) dx$$

Case 1: $b_j = 0$ and $b_h = 0$

For our application, whenever $v \to \infty$ then $b_j = 0$ and $b_h = 0$, so this is an important special case.

When $b_j = 0$ and $b_h = 0$ and $a_h = 0$ and we obtain

$$R(u, v, a_j, 0, a_h, 0) = \int_{u}^{v} a_j dx = (v - u)a_j$$

which goes to ∞ when $v \to \infty$.

When $b_j = 0$ and $b_h = 0$ and $a_h \neq 0$ and we obtain

$$\begin{aligned} R(u, v, a_j, 0, a_h, 0) &= \int_{u}^{v} a_j \exp\left(-\left[(x - u)a_h\right]\right) dx \\ &= \frac{a_j}{a_h}\left[1 - \exp\left(-\left[(v - u)a_h\right]\right)\right] \end{aligned}$$

which goes to a_j/a_h when $v \to \infty$.

Case 2: General Case with $v < \infty$

To calculate the integral, $R(u, v, a_j, b_j, a_h, b_h)$ for finite v, we can use an adaptive use of Romberg's algorithm for numeric integration (we follow closely Lange, 1999, pp. 210-211).

Let

$$f(x) = f(x, u, a_j, b_j, a_h, b_h) \quad = \quad (a_j + b_j x) \exp\left(- \left[(x-u)a_h + (x^2 - u^2)\frac{b_h}{2} \right] \right)$$

Divide the interval $[u, v]$ into n equal subintervals of length $(v-u)/n$, and let

$$T_n \quad = \quad \frac{(v-u)}{n} \left[\frac{1}{2} f(u) + \frac{1}{2} f(v) + \sum_{i=1}^{n-1} f\left(u + \frac{i(v-u)}{n} \right) \right]$$

Then $\lim_{n \to \infty} T_n = R(u, v, a_j, b_j, a_h, b_h)$.

A more accurate approximation uses Romberg's algorithm,

$$R(u, v, a_j, b_j, a_h, b_h) \quad \approx \quad \frac{4T_{2n} - T_n}{3}$$

Let $\hat{R}$ be our estimate of R. The algorithm we use to calculate $\hat{R}$ is as follows:

1. Choose n.
2. Calculate T_n.
3. Calculate T_{2n}.
4. For i=1 to I_{max} do:
   - If $|T_{2^i n} - T_{2^{i-1} n}| < \delta$ then let $\hat{R} = \frac{4T_{2^i n} - T_{2^{i-1} n}}{3}$ and stop.
   - Otherwise calculate $T_{2^{i+1} n}$, and continue.

For example, one could use $n = 100$ and $\delta = 10^{-5}$ and $I_{max} = 100$.

B Comparing the Method of Wun, Merrill, and Feuer (1998) to the PMAJ Method

Since versions of the DevCan software prior to 5.0 used the method described in Wun, Merrill, and Feuer (1998), here we compare that method to the PMAJ method. Because some calculations were slightly off in versions 5.0 and 5.1, we use a soon to be released version of DevCan with the corrected calculation. The bulk of the comparison has previously been done (see Fay et al. 2002). That comparison assumed the simple piecewise hazards models using the method described in Fay et al. (2003). The only difference between the method described in Fay et al. (2003) and that described in this paper is that in this paper we estimate the hazard functions with the PMAJ method.

In Table 3 (see pages 21-24) we recalculate Table I-15 from Ries et al. (2003) which gives lifetime risks of developing certain cancers for different race and sex combinations. We give the old method of Wun,Merrill, and Feuer (1998), the new method presented in this paper, and the percent differences. For the the Wun, Merrill, and Feuer (1998) method the age groups of the data must be in 5 year intervals except the last open ended interval. For the new method the data can be input with any age intervals, and for the example in Table 3 the first age interval is 1 year, the second is 4 years, and all subsequent intervals except the last are 5 years. Thus, the input data are slightly different for the two methods.

In general the two methods agree to within about 2 percent (see Table 3). The only cancer type with larger than about 2 percent in absolute difference is acute lymphocytic leukemia (ALL). For ALL the absolute percent differences are as large as 4.5 percent (for black males). One reason for that large absolute percent difference is the small absolute size of the ALL lifetime risk, so small absolute changes in risk translate to

large absolute percentage changes. Another reason may be that ALL is a pediatric cancer, so the differences in the input data may be part of the cause of the differences.

For age conditional probabilities of developing cancers, the methods give similar answers. Although for very small probabilities the absolute percent difference can be very large, in those cases the absolute difference is small. For large probabilities where the absolute difference between the methods may be larger, the absolute percent difference is small.

Table 3: Lifetime Risk (percent) of Being Diagnosed with Cancer by Site, Race and Sex. 12 SEER Areas, 1998-2000. (Compare to Ries, et al. 2003, Table I-15). Each cell has 3 values: PMAJ method, Wun, Merrill, and Feuer (1998) method (WMF), and percent difference= $100(PMAJ - WMF)/WMF$.

	All Races		Whites		Blacks	
Site	Males	Females	Males	Females	Males	Females
All Invasive Sites	45.33	38.77	45.52	40.07	42.75	32.27
	44.88	38.65	45.05	39.90	42.47	32.38
	1.02	0.32	1.04	0.42	0.66	-0.35
Invasive and In Situ	46.54	42.08	46.82	43.53	43.13	34.52
	46.03	41.87	46.30	43.26	42.83	34.59
	1.11	0.50	1.14	0.61	0.70	-0.20
Oral Cavity and Pharynx	1.40	0.68	1.41	0.69	1.39	0.50
	1.41	0.68	1.42	0.69	1.41	0.50
	-0.72	-0.69	-0.66	-0.66	-1.22	-0.87
Esophagus	0.76	0.26	0.77	0.25	0.77	0.38
	0.76	0.26	0.77	0.25	0.79	0.39
	-0.87	-0.79	-0.81	-0.73	-1.34	-1.23
Stomach	1.26	0.80	1.09	0.66	1.33	1.04
	1.27	0.80	1.10	0.66	1.35	1.05
	-0.77	-0.74	-0.73	-0.70	-1.13	-0.99
Colon/Rectum	5.99	5.68	6.04	5.67	4.97	5.41
	6.01	5.71	6.07	5.70	5.02	5.46
	-0.47	-0.50	-0.40	-0.43	-0.98	-0.93
Invasive and In Situ	6.33	5.95	6.38	5.93	5.33	5.74
	6.36	5.98	6.40	5.96	5.38	5.79
	-0.43	-0.48	-0.37	-0.41	-0.93	-0.90
Liver and Intrahepatic Bile Duct	0.87	0.42	0.72	0.35	0.81	0.36
	0.87	0.42	0.73	0.36	0.82	0.37
	-0.81	-0.62	-0.74	-0.49	-1.33	-1.30
Pancreas	1.24	1.24	1.23	1.22	1.28	1.35
	1.25	1.25	1.24	1.23	1.30	1.36
	-0.88	-0.83	-0.83	-0.76	-1.26	-1.21
Larynx	0.65	0.16	0.65	0.17	0.87	0.24
	0.65	0.17	0.66	0.17	0.88	0.24
	-0.85	-0.76	-0.79	-0.72	-1.32	-1.16
Invasive and In Situ	0.70	0.18	0.71	0.18	0.89	0.25
	0.71	0.18	0.71	0.18	0.91	0.25
	-0.85	-0.76	-0.79	-0.71	-1.32	-1.16

Note: Invasive cancer only unless specified otherwise

Table 3: (continued) Lifetime Risk (percent) of Being Diagnosed with Cancer by Site, Race and Sex. 12 SEER Areas, 1998-2000. (Compare to Ries, et al. 2003, Table I-15). Each cell has 3 values: PMAJ method, Wun, Merrill, and Feuer (1998) method (WMF), and percent difference= $100(PMAJ - WMF)/WMF$.

	All Races		Whites		Blacks	
Site	Males	Females	Males	Females	Males	Females
Lung and Bronchus	7.78	5.80	7.77	6.09	8.35	5.40
	7.84	5.85	7.83	6.13	8.45	5.47
	-0.78	-0.75	-0.73	-0.68	-1.20	-1.16
Melanomas of Skin	1.83	1.24	2.17	1.48	0.11	0.08
	1.84	1.24	2.18	1.49	0.11	0.08
	-0.70	-0.61	-0.61	-0.54	-1.18	-0.07
Invasive and In Situ	2.89	1.97	3.39	2.33	0.13	0.12
	2.91	1.98	3.41	2.34	0.13	0.12
	-0.63	-0.60	-0.53	-0.52	-1.20	-0.42
Breast	0.11	13.55	0.12	14.31	0.14	10.20
	0.11	13.56	0.12	14.31	0.14	10.27
	-0.74	-0.08	-0.73	0.03	-0.92	-0.70
Invasive and In Situ	0.13	16.10	0.13	16.96	0.15	12.21
	0.13	16.09	0.13	16.92	0.15	12.28
	-0.75	0.08	-0.73	0.20	-0.94	-0.57
Cervix	NA	0.79	NA	0.76	NA	0.98
	NA	0.80	NA	0.76	NA	0.99
	NA	-0.60	NA	-0.51	NA	-1.08
Corpus and Uterus, NOS	NA	2.63	NA	2.83	NA	1.68
	NA	2.65	NA	2.85	NA	1.70
	NA	-0.62	NA	-0.53	NA	-1.14
Invasive and In Situ	NA	2.67	NA	2.88	NA	1.70
	NA	2.69	NA	2.90	NA	1.72
	NA	-0.61	NA	-0.53	NA	-1.13
Ovary	NA	1.72	NA	1.85	NA	1.11
	NA	1.73	NA	1.86	NA	1.12
	NA	-0.71	NA	-0.64	NA	-1.18
Prostate	17.34	NA	16.95	NA	20.54	NA
	17.22	NA	16.83	NA	20.39	NA
	0.69	NA	0.71	NA	0.77	NA
Testis	0.35	NA	0.42	NA	0.10	NA
	0.36	NA	0.42	NA	0.10	NA
	-0.35	NA	-0.34	NA	-0.48	NA

Note: Invasive cancer only unless specified otherwise

Table 3: (continued) Lifetime Risk (percent) of Being Diagnosed with Cancer by Site, Race and Sex. 12 SEER Areas, 1998-2000. (Compare to Ries, et al. 2003, Table I-15). Each cell has 3 values: PMAJ method, Wun, Merrill, and Feuer (1998) method (WMF), and percent difference= $100(PMAJ - WMF)/WMF$.

	All Races		Whites		Blacks	
Site	Males	Females	Males	Females	Males	Females
Urinary Bladder (In Situ and Inv)	3.53	1.14	3.93	1.22	1.43	0.78
	3.55	1.14	3.94	1.23	1.45	0.79
	-0.54	-0.77	-0.44	-0.69	-1.15	-1.21
Kidney and Renal Pelvis	1.46	0.87	1.54	0.91	1.24	0.86
	1.48	0.88	1.55	0.92	1.26	0.87
	-0.78	-0.75	-0.70	-0.73	-1.33	-0.77
Brain and Other Nerv Sys	0.67	0.53	0.75	0.59	0.32	0.30
	0.67	0.53	0.75	0.59	0.33	0.31
	-0.93	-0.70	-0.82	-0.69	-1.55	-1.37
Thyroid	0.30	0.85	0.32	0.87	0.14	0.46
	0.30	0.85	0.32	0.88	0.15	0.47
	-0.74	-0.57	-0.66	-0.49	-1.30	-1.08
Hodgkin's Disease	0.23	0.20	0.26	0.22	0.19	0.15
	0.24	0.20	0.26	0.22	0.19	0.15
	-0.78	-0.59	-0.72	-0.52	-1.22	-0.95
Non-Hodgkin's Lymphomas	2.13	1.80	2.26	1.91	1.18	1.05
	2.15	1.81	2.28	1.93	1.19	1.06
	-0.78	-0.77	-0.71	-0.70	-1.26	-1.13
Myeloma	0.66	0.54	0.65	0.49	0.89	0.93
	0.67	0.54	0.66	0.50	0.91	0.94
	-0.88	-0.87	-0.82	-0.81	-1.33	-1.27
Leukemias	1.45	1.03	1.55	1.09	0.90	0.74
	1.47	1.04	1.56	1.10	0.91	0.75
	-0.90	-0.89	-0.86	-0.88	-1.40	-1.03
Acute Lymphocytic Leukemia	0.12	0.10	0.13	0.12	0.06	0.06
	0.12	0.11	0.13	0.12	0.07	0.06
	-3.39	-2.46	-3.51	-2.87	-3.85	0.88
Chronic Lymphocytic Leukemia	0.47	0.29	0.51	0.31	0.30	0.19
	0.47	0.29	0.52	0.32	0.30	0.20
	-0.77	-0.74	-0.69	-0.66	-1.32	-1.25
Acute Myeloid Leukemia	0.45	0.35	0.47	0.36	0.29	0.28
	0.45	0.36	0.47	0.37	0.29	0.28
	-0.59	-0.76	-0.56	-0.70	-0.98	-1.26

Note: Invasive cancer only unless specified otherwise

Table 3: (continued) Lifetime Risk (percent) of Being Diagnosed with Cancer by Site, Race and Sex. 12 SEER Areas, 1998-2000. (Compare to Ries, et al. 2003, Table I-15). Each cell has 3 values: PMAJ method, Wun, Merrill, and Feuer (1998) method (WMF), and percent difference= $100(PMAJ - WMF)/WMF$.

	All Races		Whites		Blacks	
Site	Males	Females	Males	Females	Males	Females
Chronic Myeloid Leukemia	0.19	0.14	0.20	0.14	0.14	0.10
	0.20	0.14	0.20	0.14	0.14	0.10
	-0.76	-0.83	-0.68	-0.78	-1.39	-1.05

Note: Invasive cancer only unless specified otherwise