Population Health Metrics

BioMed Central

Research

Open Access

# Estimating age conditional probability of developing disease from surveillance data

Michael P Fay*[1,2]

Address: [1]National Cancer Institute 6116 Executive Blvd., Suite 504 Bethesda, MD 20892-8317, USA and [2](Current Address) National Institute of Allergy and Infectious Diseases, 6700 B Rockledge Drive MSC 7609, Bethesda, MD 20892-7609, USA

Email: Michael P Fay* - faym@mail.nih.gov

* Corresponding author

Published: 27 July 2004

*Population Health Metrics* 2004, **2**:6 doi:10.1186/1478-7954-2-6

Received: 15 July 2003
Accepted: 27 July 2004

This article is available from: http://www.pophealthmetrics.com/content/2/1/6

## 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. Similarly the formula uses rates of death from disease and death from other causes per person-years alive. The rates per person-years alive are much easier to estimate than per person-years alive and disease-free. 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.

## Background

Fay, Pfeiffer, Cronin, Le, and Feuer [1] 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 [1] 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 [1] 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 [2]). Similar graphs showing the MAJ model can be made for the other rates required in the calculations, death from cancer and death from other causes per person-years alive.

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. The second model uses a series of piecewise constant values to approximate the mid-age group joinpoint model. We call this second model 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 only differs from the piecewise constant model of Fay et al [1] 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 [3] uses the PMAJ method. (There was a small calculation error in versions 5.0 and 5.1 that has been corrected in 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. The review and overview section reviews the issues in estimating the age conditional probability of developing disease from surveillance data. This section includes a motivation for using this type of statistic to describe population data. The review and overview section additionally gives graphical descriptions of the MAJ and PMAJ methods. The paper is structured so that readers not interested in the details may skip the next two sections and the appendix, which give precise and notationally involved definitions of the MAJ estimators. The examples and discussion section gives examples of the estimator of ACPDvD using three different methods for estimating the rates, the simple piecewise constant method proposed in Fay et al [1], the MAJ method, and the PMAJ method. In supplimental material [see Additional file 1] we compare the PMAJ method with the method of Wun, et al [4], since the latter method was the method used in versions of the DevCan software before version 5.0.

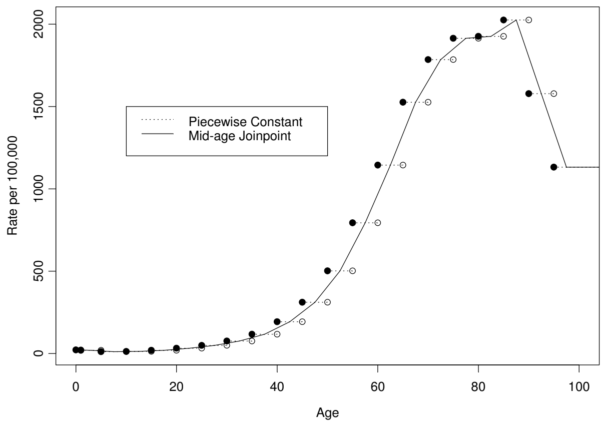

**Figure 1**
SEER 12 all invasive cancer incidence rates, 1998–2000, all races, both sexes: Piecewise constant and mid-age joinpoint methods.

## Review and overview

Consider a surveillance program like the SEER program of the U.S. National Cancer Institute. This program attempts to count every incidence of cancer within the catchment area of the program. Because cancer is a disease in which the rates of the disease are highly dependent on age, in order to give interpretability to the counts within the SEER registries, we must somehow account for the age distribution in the popoulation.

One simple and popular statistic is the age adjusted rate or directly standardized rate (DSR). In the SEER Cancer Statistics Review [2] DSRs are used to compare different cancer sites, trends on specific cancer sites over time, and rates by sex and race. The DSR is calculated by a simple weighted sum of the age specific rates for each 5 year age group, where the weights are proportional to the U.S. 2000 population. Thus, the DSR may be interpreted as the rates adjusted as if all the populations being compared had age distributions similar to the U.S. 2000 population. The DSRs are useful for gaining an overall picture of how the incidence and mortality of each cancer effects different populations (e.g., different races, SEER population at different times), while controling for the effect of differing age distributions between populations being compared. A disadvantage of the DSR is that it is hard to relate to an individual's risk. For example, Table I-4 of the SEER Cancer Statistics Review, 1975–2000 [2] states that the DSR for breast cancer for females for the years 1996–2000 is 135 per 100,000 person-years. The average American woman may wonder, how does that relate to my risk? Will I be likely to get breast cancer in my lifetime? If I am 40 years old now, what is my risk of getting breast cancer in the next 10 years given that I have survived to this old without getting it? These questions are the motivation for using the age conditional probability of developing disease (ACPDvD), and in order to estimate the ACPDvD for female breast cancer, we require information not only about the rate of female breast cancer but also about the rates of dying from female breast cancer and dying from other causes.

The ACPDvD uses cross-sectional incidence and mortality rates to estimate the age-conditional probabilities of developing disease in a hypothetical cohort in which we assume the age specific rates do not change over time. This gives a personal interpretation to the cross-sectional data, allowing statements like the following: if the incidence and mortality rates remain at their present values (as observed in SEER 12, 1998–2000), then a female born today would have a 13.5% chance of developing breast cancer over her lifetime (see Table 2). We can also calculate ACPDvD over intervals. For example, a female who has reached 40 years old without developing breast cancer has a 1.5% chance of developing breast cancer by the time she is 50.

Calculation of the ACPDvD is somewhat complicated, and we describe the complications in relation to the simple DSRs. Consider first the age specific incidence rates which are used to calculate the DSRs. These rates simply count the number of incident cases of a particular disease (e.g., female breast cancer) within each age group and divide by the total number of person-years estimated by the population. For counts of a single year, the person-years are estimated by the mid-year population of the catchment area (for sex-specific cancers like prostate cancer or female breast cancer, we only use the population of the appropriate sex). Note that the incident cases may include individuals who have previously been diagnosed with the cancer and have developed a new primary cancer.

For the ACPDvD for any specific disease we would like the rate of first incidence per person-years alive and disease-free. Thus, there are two difficulties, (1) the usual age specific incidence rates include persons with multiple primary cancers, and (2) the denominators include persons who have previously been diagnosed. Merrill and Feuer [5] discuss both difficulties and adjust for them creating risk-adjusted cancer incidence rates. Merrill and Feuer [5] study the effect of these adjustments for several cancer sites. To handle the first difficulty, (similar to [5]) we can remove cases where we have a record of a previous diagnosis of that particular type of cancer. Because the registries in SEER were not all begun at the same time, to avoid bias the DevCan program only searches the records for previous cancers back until the year when the last registry was added. This year is denoted the follow-back year. (If the disease of interest is any malignant cancer, then the difficulty is handled differently. Although at each cancer record we do not record what specific types of cancers were previously diagnosed for the person, we do know whether any tumors were previously diagnosed. Thus, if the disease of interest is any malignant cancer and if the record states there was a previously diagnosed tumor, then we assume that the previously diagnosed tumor was malignant, and do not count that case as a first incidence.) To handle the second difficulty, the additional person-years in the denominator, Merrill and Feuer [5] adjust the denominator by multiplying the age-specific population by 1 minus an estimate of the prevalence of the disease in the population. Merrill and Feuer [5] also estimate the prevalence of medical procedures which remove individuals from the at-risk population, such as hysterectomy which removes the risk of uterine cancers.

In calculating the ACPDvD we use only first incident of the disease of interest as in [5], but we correct for the denominators in a different way using an assumption and some mathematics from the theory of competing risks. This second correction is detailed with precise mathematical notation in Fay et al [1]; here we give more heuristic arguments.

In the following let the disease of interest be "cancer". The ACPDvD between ages $x$ and $y$, given alive and cancer-free at age $x$, may be written as the fraction,

$$\frac{\text{Probability that the first cancer occurred between the ages of } x \text{ and } y}{\text{Probability alive and cancer-free at age } x}.$$

To calculate the numerator, we integrate over the probability that the first cancer occurred at exactly age a. In math notation this probability is

$$\int_x^y f_c(a)\,da$$

where $f_c(a)$ is a probability function representing the probability that the first cancer occurred at exactly age a.

**Table 1: Notation**

Random Variables and Parameters

$T$ = age at death
$J$ = type of death
($J = d$) = death from cancer
($J = o$) = death from other causes
$\lambda_c(t)$ = rate at $t$ for first cancer given alive

$\lambda_o(t)$ = rate at $t$ for death before cancer given alive

$\lambda_d(t)$ = rate at $t$ for death from cancer given alive
$\lambda_a(t)$ = rate at $t$ for death given alive

$$S_j(t) = \exp\left\{-\int_0^t \lambda_j(u)\,du\right\}$$
$$\text{for } j = a,c,o,d$$

$T^*$ = age at first cancer or death before cancer
$J^*$ = type of event
($J^* = c$) = first cancer
($J^* = o$) = death before first cancer

$\lambda_c^*(t)$ = rate at $t$ for first cancer given alive and cancer-free

$\lambda_o^*(t)$ = rate at $t$ for death before cancer given alive and cancer-free

$\lambda_a^*(t)$ = rate at $t$ for first cancer or death before first cancer given alive and cancer-free

$$S_j^*(t) = \exp\left\{-\int_0^t \lambda_j^*(u)\,du\right\}$$
$$\text{for } j = a,c,o$$

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

$d_i$ = number of cancer deaths
$o_i$ = number of other deaths

$n_i^{(j)}$ = 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)

**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) | Prostat(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 |

One key result described in Fay et al [1] is that $f_c(a)$ can be written as the product of two functions,

$\lambda_c(a)$ = the probability that the first cancer occurred at exactly age a, given the individual is alive just before age $a$, and

$S_a(a\text{-})$ = the probability that the individual is alive just before age $a$.

The function $\lambda_c(a)$ is known as a cause-specific hazard function, and it is estimated by some function of the age-specific rates, such as the piecewise constant model of Fay et al [1] or the MAJ model introduced in this paper (see Figure 1). Using standard results for continuous survival data, we can write $S_a(a\text{-})$ as

$$S_a(a-) = \exp\left(-\int_0^a \lambda_a(u)\,du\right.$$

where $\lambda_a(u)$ ( = the probability that the individual died at age $u$, given the individual is alive just before age $u$) is the usual hazard function. We estimate $\lambda_a(u)$ using some function of the age-specific rates. Thus, the numerator can be written as

$$\int_x^y \lambda_c(a)\exp\left(-\int_0^a \lambda_a(u)\,du\right)da$$

If we use the MAJ for both hazard functions, then there is no closed form solution. To see why this is so, note that within the exponential, the integral of a piecewise linear function is the sum of a series of quadratic functions, and the overall integral has no closed form solution. This problem motivates the piecewise mid-age joinpoint (PMAJ) model, where we use a series of piecewise constant functions to approximate the MAJ model. Figure 2 gives the PMAJ model together with the piecewise constant model used by Fay et al [1] for 70 to 90 year olds from the SEER 12, 1998–2000 rates for all invasive (first) cancer incidence rates per person-years alive. Remember, although both Figure 1 and Figure 2 plot incidence rates, we additionally need similar rate functions for mortality rates to calculate the ACPDvD.

Now consider the denominator of the ACPDvD, the probability of being alive and cancer-free at age $x$, denoted $S_a^*(x)$. For reference, in Table 1 we give the notation. The only change from the notation in Fay et al [1] 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. Fay et al [1] assumed that the risk of death from other causes does not change if you have previously been diagnosed with cancer, then used the key result mentioned above together with some algebra and calculus to derive the denominator. Then the ACPDvD between the ages of $x$ and $y$ given alive and cancer-free just before age $x$ is

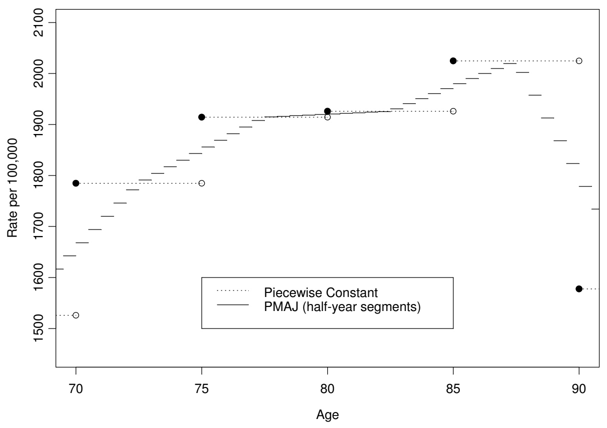

**Figure 2**
SEER 12 all invasive cancer incidence rates, 1998–2000, all races, both sexes: Piecewise constant and PMAJ methods.

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

The details of the MAJ and the PMAJ models are given in the next two sections.

Readers only interested in the practical ramifications of the choice in models may skip to the examples and discussion section.

## Mid-age group joinpoint estimator

In Fay et al [1], 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 [1] 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 [1].) The MAJ estimator for the rate of event $j$ (for $j$ = c, d, or $o$) at $t_i$ (for $i = 0,1,..., k$) is

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

where $j_i$ is either $c_i$, $d_i$, or $o_i$ as defined in Table 1. (Note that $\tilde{\lambda}_j\left(t_i\right) = \hat{\lambda}_j\left(a_i\right) = \hat{\lambda}_j\left(t_i\right)$, where $\hat{\lambda}_j\left(\cdot\right)$ is the piecewise constant function used in [1]). 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

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

Then for t ∈ $[t_i, t_{i+1})$ for $i = 1,..., k$, we define $\tilde{\lambda}_j\left(t\right)$ 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,

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

Where

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

$$(5)$$

and

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

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

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

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

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

so that for $i = 0,1,...,$k,

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

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

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

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,

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

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,..., $i$ and $v \leq t\ell_{+1}$) is defined implicitly (see the Appendix). Then,

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

### 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.

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

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 $i$th interval ($i$ = 1,...,$k$) for the $j$th 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 $i$th 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 [3]), 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 $h$th (for $h$ = 1,..., $m_i$) of the $m_i$ intervals within interval $i$ (for $i$ = 1,...,$k$) 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

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

(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

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

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 [1] 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 $i = 0,1,\dots, k$, 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 [1]. 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 $b_0 = \tau_0,\dots, b_i = \tau_i$, $b_{i+1} = x$, $b_{i+2} = \tau_{i+1},\dots, b_{j+1} = \tau_j$, $b_{j+2} = y$, $b_{j+3} = \tau_{j+1},\dots, b_{M+3} = \tau_{M+1} = \infty$. We let

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

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 $\phi(\lambda,\ell) = \int_{b_\ell}^{b_{\ell+1}} \exp(-(u-b_\ell)\lambda)du$. These integrals are

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

where the case $\lambda = 0$ and $b\ell_{+1} = \infty$ is one of the "impossible" hypothetical cohorts (see Section 3.1 of [1]). Thus, we obtain,

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

## Examples and discussion

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

For all of the examples we use data from 1998–2000 [6]. 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 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 ACP-DvD, and we have shown that the PMAJ method provides fast and reasonable estimators for the rates.

## Appendix: 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 $\alpha_j\ell = \alpha_j, \beta_j\ell = b_j, \alpha_h\ell = a_h$, 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 $\infty$ 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 [7], pp. 210–211).

Let

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

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

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

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,

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

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 $\left|T_{2^i n} - T_{2^{i-1} n}\right| < \delta$ then let $\hat{R} = \dfrac{4T_{2^i n} - T_{2^{i-1} n}}{3}$ and stop.

- Otherwise calculate $T_{2^{i+1} n}$, and continue.

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

## Additional material

### Additional File 1

*Comparing the method of Wun, Merrill, and Feuer (1998) to the PMAJ method. We calculate lifetime risks of developing certain cancers for different race and sex combinations. For each lifetime risk we give the old method of Wun, Merrill, and Feuer [4], the PMAJ method, and the percent difference. In general, the two methods agree to within about 2 percent.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1478-7954-2-6-S1.pdf]

## 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

1. Fay MP, Pfeiffer R, Cronin KA, Le C, Feuer EJ: **Age-conditional probabilities of developing cancer.** *Statistics in Medicine* 2003, **22(11):**1837-1848.
2. 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* 2003 [http://seer.cancer.gov/csr/1975_2000].
3. DevCan: **Probability of Developing or Dying of Cancer Software, Version 5.2.** *Statistical Research and Applications Branch, National Cancer Institute* 2004 [http://srab.cancer.gov/devcan].
4. Wum L-M, Merrill RM, Feuer EJ: **Estimating lifetime and age-conditional probabilities of developing cancer.** *Lifetime Data Analysis* 1998, **4:**169-186.
5. Merrill RM, Feuer EJ: **Risk-adjusted cancer-incidence rates (United States).** *Cancer Causes and Control* 1996, **7:**544-552.
6. . Surveillance, Epidemiology, and End Results (SEER) Program http://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 sub mission. Underlying mortality data provided by NCHS http://www.cdc.gov/nchs.
7. Lange K: *Numerical Analysis for Statisticians* Springer:New York; 1999.