Lifetime Data Analysis, 4, 169–186 (1998)
© 1998 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Estimating Lifetime and Age-Conditional Probabilities of Developing Cancer

LAP-MING WUN
Cancer Control Research Program, National Cancer Institute, Applied Research Branch, Bethesda, MD, U.S.A.

RAY M. MERRILL
Cancer Control Research Program, National Cancer Institute, Applied Research Branch, Bethesda, MD, U.S.A.

ERIC J. FEUER
Cancer Control Research Program, National Cancer Institute, Applied Research Branch, Bethesda, MD, U.S.A.

Received June 26, 1996; Revised November 17, 1997; Accepted December 1, 1997

Abstract. Lifetime and age-conditional risk estimates of developing cancer provide a useful summary to the public of the current cancer risk and how this risk compares with earlier periods and among select subgroups of society. These reported estimates, commonly quoted in the popular press, have the potential to promote early detection efforts, to increase cancer awareness, and to serve as an aid in study planning. However, they can also be easily misunderstood and frightening to the general public.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute and the American Cancer Society have recently begun including in annual reports lifetime and age-conditional risk estimates of developing cancer. These risk estimates are based on incidence rates that reflect new cases of the cancer in a population free of the cancer. To compute these estimates involves a cancer prevalence adjustment that is computed cross-sectionally from current incidence and mortality data derived within a multiple decrement life table. This paper presents a detailed description of the methodology for deriving lifetime and age-conditional risk estimates of developing cancer. In addition, an extension is made which, using a triple decrement life table, adjusts for a surgical procedure that removes individuals from the risk of developing a given cancer. Two important results which provide insights into the basic methodology are included in the discussion. First, the lifetime risk estimate does not depend on the cancer prevalence adjustment, although this is not the case for age-conditional risk estimates. Second, the lifetime risk estimate is always smaller when it is corrected for a surgical procedure that takes people out of the risk pool to develop the cancer. The methodology is applied to corpus and uterus NOS cancers, with a correction made for hysterectomy prevalence. The interpretation and limitations of risk estimates are also discussed.

Keywords: Life table; incidence; cancer prevalence; hysterectomy prevalence; corpus and uterus NOS cancers.

1. Introduction

Cancer risk estimates, commonly quoted in the popular press, have the potential to promote greater cancer awareness, early detection efforts, and may be useful in study planning (e.g., Schwab, 1991; One in nine, 1992). However, they can also be easily misunderstood and create an unrealistically frightening perception of risk (Blakeslee, 1992). Thus, it is important that the statistical community understand the derivation of these risk estimates so that these statistics can be properly computed and interpreted for the general public.

Risk estimates of developing cancer are derived utilizing a three state stochastic process where the transient state represents persons alive and cancer free, and two absorbing states represent developing cancer and death from other causes in the absence of cancer. The lifetime risk of developing cancer represents the ultimate probability of being absorbed in the cancer state. The age-conditional risk of developing cancer from age x to $x+k$ represents the probability of being absorbed in the cancer state during this interval, conditional on not being absorbed prior to age x. This stochastic process is represented using a multiple decrement life table.

The derivation of the life table used to estimate lifetime and age-conditional risk estimates of disease has a long history (Goldberg et al., 1956; Zdeb, 1977; Seidman et al., 1978; and Kramer et al., 1980). However, criticism of the standard lifetime risk methodology has arisen when based on incidence rates that are derived using more than just the first instance of a given cancer and which reflect rates among the total rather than the population free of the cancer (Bender et al., 1992). Standard reports of incidence rates (e.g., Ries et al., 1997) do not adjust for either of these two factors. The methodology presented by Feuer et al. (1993) for deriving lifetime and age-conditional risk estimates addresses these concerns, and is currently used to derive risk estimates annually reported by the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (Ries et al., 1997) and the American Cancer Society (Parker et al., 1996). This methodology has been accompanied by software for estimating the probability of developing cancer (Feuer and Wun, 1994).

The purpose of this paper is to give a more mathematical description of the Feuer et al. (1993) methodology, along with an extension which considers the case where a surgical procedure removes a person from the cancer risk population (i.e., a third absorbing state is added to the process). Two new results which provide insights into the basic methodology are included in the discussion. First, the lifetime risk estimate does not depend on the cancer prevalence adjustment, although this is not the case for age-conditional risk estimates. Second, the lifetime risk estimate is always smaller when it is corrected for a surgical procedure that takes people out of the risk pool to develop the cancer. Interpretation and limitations of lifetime and age-conditional risk estimates are also considered.

2. Methods

Cross-sectional incidence, mortality, and population vital statistics data are required to derive lifetime and age-conditional risk estimates of developing cancer. The incident cases consist of only the first occurrence of a given cancer; the mortality cases consist of deaths attributed to cancer and non-cancer causes; and the population consists of the mid-year population, which is assumed to represent the person-years at risk in a cohort of individuals aging through successive five-year age intervals. Incidence and mortality rates obtained from this data are converted to probabilities using an exponential model, and applied to a hypothetical cohort of 10 million live births. Estimates are then derived for each five-year age group in the hypothetical cohort of the number alive and cancer free at the beginning of the interval; the number of newly developed cancers in the interval; and the number of non-cancer deaths in the interval among the cancer free population. Lifetime and age-conditional probabilities of developing cancer are also computed. Herein the term "cancer" refers to a specified cancer of interest and the term "develop cancer" refers to diagnosed cancers only, and does not include cancers that have not yet been detected.

Exponential survival with a constant hazard is assumed in each age interval (i.e., piecewise exponential survival), where the probability of an event in an interval of length i is

$$1 - \exp(-i\lambda), \tag{1}$$

where λ is the rate of the event. Although many life table calculations are constructed utilizing the assumption that the probability of an event occurs in the interval which follows an uniform distribution, we utilize the exponential assumption because calculations for the last open-ended age interval become tractable, with results differing only slightly from the uniform in the earlier intervals.

2.1. Notation

Notation used in the computations of lifetime and age-conditional risk estimates of developing cancer are presented in Table 1. Capital letters are used to represent observed vital statistics data. Lower case letters are used to represent derived quantities computed by applying the vital statistics data to the hypothetical cohort.

Table 1. Letter definitions.

Letter		Definition
Vital Statistics Data		
L	=	mid-year population is assumed to represent the average population at risk of a cohort of individuals aging through the interval (e.g., if deaths are assumed to occur uniformly over a one year time period, the mid-year population for a 5 year age interval can be interpreted as the person-years at risk of a single cohort observed for 5 years);
C	=	number of first occurrences of cancer of a certain type; and
D	=	number of deaths.
Derived Quantities		
l	=	population at the beginning of the interval in the hypothetical cohort;
p	=	person years at risk during the interval in the hypothetical cohort;
d	=	number of non-cancer deaths in the interval among those cancer free when they die;
a	=	number of newly developed cancers in the interval;
r	=	incidence rate;
g	=	probability of developing cancer;
m	=	death rate; and
q	=	probability of dying in the interval.

Subscripts and superscripts are defined following the notation of Seidman et al. (1978). The left subscript denotes the number of years in the interval, if relevant. The right subscript denotes the age at the start of the interval. Superscripts denote cancer status, as defined in Table 2.

Table 2. Superscript definitions.

Superscript	Definition
The left superscript denotes the following:	
0 -	persons cancer free at the beginning of the interval, and
Blank -	total persons at the start of the interval, if relevant.
The right superscript denotes the following:	
0 -	died due to causes other than cancer, and
Blank -	total who died, if relevant.

2.2. *Computation*

The computational steps presented in this section are applied to nineteen five-year age intervals 0–4, 5–9, ..., 90–94 and, using a modification of these steps, the final open interval 95+. The number of cancer free survivors at age x is based on a hypothetical cohort of $l_0 = 10$ million live births. In the first interval, because everyone is cancer free at birth, ${}^0l_0 = l_0$. The number alive and cancer free at the beginning of successive age intervals is denoted by ${}^0l_{x+5}$; the number of newly developed cancers through the interval is denoted by ${}_5a_x$; and the number of non-cancer deaths in the interval among the cancer free population is denoted by ${}_5d_x$. Based on the standard results from the theory of competing risks (Gail, 1975):

$$ {}^0l_{x+5} = ({}^0l_x)(\exp(-5({}^0_5m^0_x + {}^0_5r_x))), \tag{2} $$

$$ {}_5a_x = ({}^0l_x)(1 - \exp(-5({}^0_5m^0_x + {}^0_5r_x)))({}^0_5r_x/({}^0_5m^0_x + {}^0_5r_x)), \tag{3} $$

and

$$ {}_5d_x = ({}^0l_x)(1 - \exp(-5({}^0_5m^0_x + {}^0_5r_x)))({}^0_5m^0_x/({}^0_5m^0_x + {}^0_5r_x)), \tag{4} $$

where ${}^0_5m^0_x$ is the non-cancer related death rate among persons cancer free at the beginning of the five-year age interval, and ${}^0_5r_x$ is the incidence rate among persons cancer free at the beginning of the five-year age interval. These equations apply to the first nineteen five-year age intervals. For the last open ended age interval, the estimates are computed:

$$ \begin{aligned} a_{95+} &= {}^0l_{95+}\{1 - \exp[-\infty({}^0r_{95+} + {}^0m^0_{95+})]\}[{}^0r_{95+}/({}^0r_{95+} + {}^0m^0_{95+})] \\ &= {}^0l_{95+}({}^0r_{95+}/({}^0r_{95+} + {}^0m^0_{95+})) \end{aligned} \tag{5} $$

and

$$ d_{95+} = {}^0l_{95+}({}^0m^0_{95+}/({}^0r_{95+} + {}^0m^0_{95+})). \tag{6} $$

The lifetime probability of developing cancer from birth through age $x + 5$, where x (i.e.,

$x = 0, 5, \ldots, 90, 95+$) indexes the age at the beginning of the interval, is

$$\sum_{i=0}^{x/5} {}_5a_{5i} \Big/ {}^0l_0. \tag{7}$$

The probability of developing cancer by age $x+5$, given that the person is currently cancer free at age j, is

$$\sum_{i=j/5}^{x/5} {}_5a_{5i} \Big/ {}^0l_j, \tag{8}$$

where $j = 0, 5, \ldots, 90, 95+$.

2.3. *Derivation of Key Quantities in the Computation of ${}^0l_{x+5}$, ${}_5a_x$, and ${}_5d_x$*

We begin by assuming that the following data is available in a specified population based catchment area: the number of first occurrences of cancer of a certain type, ${}_5C_x$; the mid-year population, ${}_5L_x$; the total number of deaths, ${}_5D_x$; and the number of deaths due to non-cancer causes, ${}_5D_x^0$. The cancer incidence rate among the cancer free population, ${}_5^0r_x$, cannot be calculated directly because the number cancer free, ${}_5^0L_x$, is not usually available. Instead, we first derive the probability of developing cancer among the total population, ${}_5g_x$, from the incidence rate, ${}_5r_x$, through equation (1).

$$\begin{aligned} {}_5r_x &= \frac{{}_5C_x}{{}_5L_x} \\ {}_5g_x &= 1 - \exp(-5 \times {}_5r_x) \end{aligned} \tag{9}$$

Because the first occurrences of cancer come only from the cancer free population,

$${}_5^0g_x = {}_5g_x \left[\frac{l_x}{{}^0l_x} \right], \tag{10}$$

where $l_x/{}^0l_x$ can be thought of as a prevalence correction factor. Replacing ${}_5g_x$ with ${}_5^0g_x$ and ${}_5r_x$ with ${}_5^0r_x$ in equation (9), and solving for the cancer free incidence rate gives

$${}_5^0r_x = (-1/5)\log(1 - {}_5^0g_x).$$

Assuming the birth and death rates are constant over calendar time, the death rate can be computed directly from the data as

$${}_5m_x = \frac{{}_5D_x}{{}_5L_x}. \tag{11}$$

The death rate is converted into the probability of death through (1); that is,

$${}_5q_x = 1 - \exp(-5 \times {}_5m_x).$$

The number of persons at the beginning of successive five-year age intervals is computed as

$$l_{x+5} = l_x(1 - {}_5q_x).$$

The death rate due to non-cancer causes ${}_5m_x^0$ is also computed directly from the data

$${}_5m_x^0 = \frac{{}_5D_x^0}{{}_5L_x}.$$

Thus the prevalence correction factor is recursively defined, being known in the first interval (i.e., $l_0/{}^0l_0 = 1$) with ${}_5g_x$ computed from ${}_5r_x$ and ${}_5^0g_x$, which is required to derive ${}_5^0r_x$, obtained using the correction factor. The values ${}_5m_x$ and ${}_5m_x^0$ are derived directly from the data. In the second and later intervals, the prevalence correction factor (i.e., $l_x/{}^0l_x$) is a function of values in the preceding interval.

Because death rates are generally not available for the site-specific cancer free population, we assume that the rate of non-cancer death is the same in the total population as in the cancer free population (i.e., ${}_5^0m_5^0 = {}_5m_x^0$). This may be a reasonable assumption because the total population is comprised mostly of the cancer free population. (Relative survival rates, as annually reported by the National Cancer Institute, make the even stronger assumption that mortality rates for the total population can be applied to the cancer population. For some cancers this is a reasonable approximation, whereas for others (e.g., lung cancer), the assumption is questionable (Brown et al., 1993).) Hence, ${}_5^0q_x^0 = {}_5q_x^0$ because

$${}_5^0q_x^0 = 1 - \exp(-5 \times {}_5^0m_x^0) \quad \text{and} \quad {}_5q_x^0 = 1 - \exp(-5 \times {}_5m_x^0).$$

To compute a_{95+} and d_{95+} in equations (5) and (6) require some modifications. While ${}^0l_{95+}$ and ${}^0m_{95+}^0$ are computed in the same manner as in the younger age intervals, ${}^0r_{95+}$ cannot be, but is computed as follows:

$${}^0r_{95+} = (\omega/1 - \omega)){}^0m_x^0 \tag{12}$$

where

$$\omega = [l_{95+}/(m_{95+} \times {}^0l_{95+})]r_{95+}$$

The derivation of equation (12) is shown in Appendix A, Part I.

3. Extended Method: Correction for a Surgical Procedure

An additional adjustment to the total population is required for certain cancers to correct for prevalent cases of a surgical procedure that removes individuals from risk of developing the cancer. For example, the development of uterine cancers is influenced by the fact that many women are not at risk for this cancer because of a prior hysterectomy. We focus on hysterectomy, although other surgical procedures could be considered (e.g., cholocystectomy, due to cholelithiasis gallstones).

3.1. Extended Notation

Notation used in the computations of lifetime and age-conditional risk estimates of developing cancer are extended to incorporate a hysterectomy correction. Additional letters and modified superscript definitions specific to analyses for uterine cancer are presented in Table 3.

Table 3. Hysterectomy and superscript definitions.

Letter and Superscript Symbols		Definition
Vital Statistics Data		
H	=	number of hysterectomies; and
C^h	=	number of new cancer cases treated by a hysterectomy.
Derived Quantities		
h	=	hysterectomy rate;
h^c	=	hysterectomy rate among diagnosed cancer cases;
f	=	probability of having hysterectomy;
s	=	number of hysterectomies performed among those cancer free at the time of surgery; and
a^{sp}	=	number of newly developed cancers in the interval, corrected for hysterectomy.
Superscripts		
The left superscript denotes the following:		
	0 -	persons cancer free at the beginning of the interval, and
	Blank -	total persons at the start of the interval, if relevant.
The right superscript denotes the following:		
	0 -	died or had a hysterectomy due to causes other than cancer, and
	Blank -	total who died or had a hysterectomy, if relevant.

3.2. Extended Computation

A triple decrement life table is used to subtract out the number who develop cancer, have a hysterectomy (for non-cancer causes), or die of other causes than cancer. This process yields the cancer and hysterectomy free population at the start of each age interval. In the case of a hysterectomy correction, equations (2) through (4) are modified to include the hysterectomy rate among women cancer and hysterectomy free at the beginning of the five-year age interval, who had a hysterectomy for non-cancer causes, ${}^0_5h^0_x$. Again, using the standard results from the theory of competing risks (Gail, 1975):

$${}^0l_{x+5} = ({}^0l_x)(\exp(-5({}^0_5m^0_x + {}^0_5r_x + {}^0_5h^0_x))),$$

$${}_5a_x = ({}^0l_x)(1 - \exp(-5({}^0_5m^0_x + {}^0_5r_x + {}^0_5h^0_x)))({}^0_5r_x/({}^0_5m^0_x + {}^0_5r_x + {}^0_5h^0_x)),$$

and

$$ {}_5d_x = ({}^0l_x)(1 - \exp(-5({}^0_5m^0_x + {}^0_5r_x + {}^0_5h^0_x)))({}^0_5m^0_x/({}^0_5m^0_x + {}^0_5r_x + {}^0_5h^0_x)). $$

In addition, the number of hysterectomies performed among those cancer free at the time of diagnosis is computed as

$$ {}_5s_x = ({}^0l_x)(1 - \exp(-5({}^0_5m^0_x + {}^0_5r_x + {}^0_5h^0_x)))({}^0_5h^0_x/({}^0_5m^0_x + {}^0_5r_x + {}^0_5h^0_x)). $$

These equations apply to the first nineteen five-year age intervals. For the last open ended age interval

$$ \begin{aligned} a_{95+} &= {}^0l_{95+}\{1 - \exp[-\infty({}^0r_{95+} + {}^0m^0_{95+} + {}^0h^0_{95+})]\}[{}^0r_{95+}/({}^0r_{95+} + {}^0m^0_{95+} + {}^0h^0_{95+})] \\ &= {}^0l_{95+}({}^0r_{95+}/({}^0r_{95+} + {}^0m^0_{95+} + {}^0h^0_{95+})), \end{aligned} \tag{13} $$

$$ d_{95+} = {}^0l_{95+}({}^0m^0_{95+}/({}^0r_{95-} + {}^0m^0_{95+} + {}^0h^0_{95+})), \tag{14} $$

and

$$ s_{95+} = {}^0l_{95+}({}^0_5h^0_{95+}/({}^0r_{95-} + {}^0m^0_{95+} + {}^0h^0_{95+})). \tag{15} $$

For those cancers influenced by hysterectomy, a women who is cancer free is considered at risk of developing cancer, of dying from non-cancer causes, and of having a hysterectomy for non-cancer causes. As before, we assume that the following data is available: the number of new cases of cancer, ${}_5C_x$; the mid-year population, ${}_5L_x$; the total number of deaths, ${}_5D_x$; and the number of deaths due to non-cancer causes, ${}_5D^0_x$. We also assume that data is available on the number of new hysterectomies performed, ${}_5H_x$, and the number of new cases of cancer which are treated by a hysterectomy, ${}_5C^h_x$. The derivation of ${}^0_5h^0_x$ involves four steps. First, the hysterectomy rate was computed as

$$ {}_5h_x = \frac{{}_5H_x}{{}_5L_x}, $$

and the hysterectomy rate due to diagnosed uterine cancer was computed as

$$ {}_5h^c_x = \frac{{}_5C^h_x}{{}_5L_x}. $$

Second, to determine the rate of hysterectomies for non-cancer causes,

$$ {}_5h^0_x = {}_5h_x{-}{}_5h^c_x. $$

Third, the probability of having a hysterectomy for non-cancer causes was computed with ${}_5h^0_x$ through equation (1), as

$$ {}_5f_x = 1 - \exp(-5 \times {}_5h^0_x). \tag{16} $$

Fourth, we adjust ${}_5f_x$ by a prevalence correction factor $l_x/{}^0l_x$, derived recursively for each successive age interval in the life table, in order to obtain ${}^0_5f_x$; that is

$$ {}^0_5f_x = {}_5f_x \left[\frac{l_x}{{}^0l_x} \right] . \tag{17} $$

Replacing ${}_5f_x$ with ${}^0_5f_x$ and ${}_5h^0_x$ with ${}^0_5h^0_x$ in equation (16), and solving for the cancer free hysterectomy rate gives

$$ {}^0_5h^0_x = (-1/5)\log(1 - {}^0_5f_x). $$

To compute a_{95+}, d_{95+}, and s_{95+} in equations (13) through (15) requires modifications, such that ${}^0r_{95+}$ and ${}^0h^0_{95+}$ are

$$ {}^or_{95+} = [\omega/(1 - \omega - \Psi)]{}^0m^0_{95+}, \tag{18} $$

and

$$ {}^oh^0_{95+} = [\Psi/(1 - \omega - \Psi)]{}^0m^0_{95+}, \tag{19} $$

where

$$ \omega = (l_{95+} \times r_{95+})/(m_{95+} \times {}^0l_{95+}), \quad \Psi = (l_{95+} \times h^0_{95+})/(m_{95+} \times {}^0l_{95+}). $$

The derivation of equations (18) and (19) are given in Appendix A, Part II.

4. Example—Corpus and Uterus NOS Cancers

We apply this methodology using corpus and uterus NOS cancers incidence data from the nine standard registries of the National Cancer Institute's SEER Program (San Francisco-Oakland, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico, Utah, Seattle (Puget Sound), and Metropolitan Atlanta), from 1990 to 1992. We also use the National Center for Health Statistics' mortality data and the Bureau of the Census' population data, both taken from the SEER geographic areas for the same time period. The method used for obtaining only the first occurrence of a given cancer type is described elsewhere (Merrill and Feuer, 1996). Hysterectomy data was obtained from the National Hospital Discharge Survey (Graves, 1992). U.S. hysterectomy rates from 1990 to 1992 for all races combined were used to derive lifetime risk because the rates by race and region were not available. We assume the U.S. hysterectomy rates are representative of the SEER area hysterectomy rates.

A large portion of the female population is not at risk of developing corpus and uterus NOS cancers because they have had their uterus removed by a hysterectomy. Non-preganancy related hysterectomy is the most frequently performed surgical procedure in U.S. women; about one-third of women receive a hysterectomy by age 65 (Pokras, 1989). Hence, it is important to adjust for hysterectomy so the population reflects those at risk of developing corpus and uterus NOS cancers (Lyon and Gardner, 1977). A modification of the standard software used to compute lifetime and age-conditional probability estimates (i.e., the National Cancer Institute's program DEVCAN (Feuer and Wun, 1994)) allows us to estimate lifetime and age-conditional risks of developing cancer adjusted by the prevalence of hysterectomies, as derived in the life table by the cross-sectional hysterectomy rates.

Table 4 summarizes the probability of developing in situ and invasive corpus and uterus NOS cancers for all races combined, from 1990 to 1992. The number who develop the cancer peaks in the 70–74 age group. The hysterectomy rate peaks in the 45–49 age group (data not shown), whereas the number having a hysterectomy peaks in the 70–74 age group because of the larger number at risk in this age group. The number of hysterectomies in this hypothetical cohort is sufficiently large to substantially affect the size of the population at risk. With the exception for a relatively large number of infant deaths, the number that die of other causes rises through age group 85–89, and then falls. The lifetime probability of developing cancer from birth represents the cumulative number of developed corpus and uterus NOS cancer cases divided by the hypothetical population of 10 million live births. At age 70, over half that lifetime risk is realized. The lifetime risk of developing corpus and uterus NOS cancers is 2.34 percent, or 1 in 43.

The percent developing in situ and invasive corpus and uterus NOS cancers before a specified age (Z), given they are cancer free at current age (Y), is reported in Table 5. This age-conditional risk estimate represents the cumulative number of developed cases divided by the total population at the beginning of the age interval. For example, the risk from 50 to 60 is 0.6 percent, the risk from 60 to 70 is 1.1 percent, and from 70 to 80 is 1.27 percent. These age-conditional risks provide more relevant information than do lifetime risk estimates for a women at an older age when corpus and uterine cancer rates rise significantly.

5. Discussion

This article presents a complete derivation of the methodology for deriving lifetime and age-conditional risk estimates of developing cancer. These risk estimates are calculated for persons free of the specific cancer at the beginning of the age interval, and reflect the risks that prevail in the current population. They do not take into account individual behaviors and risk factors, but are based on population-based data. Risk estimates for individuals with specific risk profiles have been considered elsewhere (Gail et al., 1986). Conventionally reported incidence rates in the US reflect, in some sense, 'risk,' but the numerator in these rate calculations may include multiple diagnosed cases of the cancer in the same person and the population may include persons not at risk of developing the disease (e.g., those already diagnosed with the cancer or those not at risk of developing the disease because they have had the organ in question removed (Merrill and Feuer, 1996). In addition, an extension is developed which, using a triple decrement life table, adjusts for a surgical procedure that removes individuals from the risk of developing a given cancer. The primary aim of the extended methodology is to provide lifetime and age-conditional risk estimates based on the cancer free, at risk population.

Two results follow from the adjustment for cancer prevalence and the adjustment for a surgical procedure that removes people from the cancer risk pool. The first is that the

Table 4. Probability of developing invasive corpus and uterus NOS cancers in the standard SEER areas, 1990–92.

(1) AGE[a]	(2) TOTAL ALIVE & AT RISK AT BEGINNING OF INTERVAL[a]	(3) NUMBER WHO DEVELOP CANCER THIS INTERVAL[a]	(4) NUMBER HAVING A HYSTERECTOMY THIS INTERVAL NOT DUE TO CANCER[a,b]	(5) NUMBER WHO DIE OF OTHER CAUSES THIS INTERVAL[a,b]	(6) CUMULATIVE PROBABILITY OF DEVELOPING CANCER FROM BIRTH[a,c]
0	10000000	0	0	92716	0
5	9907284	0	0	8578	0
10	9898705	0	0	8601	0
15	9890104	21	2984	21175	2.11746E-06
20	9865925	38	40096	23337	5.87046E-06
25	9802455	388	168696	28245	4.47106E-05
30	9605125	887	282244	36939	0.000133369
35	9285055	2743	463411	48491	0.000407667
40	8770410	5330	624976	64938	0.000940621
45	8075166	9631	585646	92594	0.001903707
50	7387296	17642	281321	138737	0.00366792
55	6949595	26995	152626	214100	0.006367372
60	6555875	34577	140544	306336	0.00982506
65	6074418	38338	124980	431428	0.013658882
70	5479671	38665	106935	583474	0.017525338
75	4750597	30784	61493	776641	0.020603765
80	3881679	17519	45547	999328	0.02235567
85	2819285	7685	17441	1114914	0.023124155
90	1679245	2400	3338	977531	0.023364122
95+	695976	734	8	695234	0.023437571

[a] The symbols representing these columns are: (1) x, the beginning of each age interval; (2) ${}^{0}l_x$; (3) ${}_5a_x$; (4) ${}_5s_x$; (5) ${}_5d_x$; and (6) as given in Equation 7.

[b] Among the cancer-free population at the beginning of the interval who did not develop cancer during the interval.

[c] Applies through the end of the interval.

Table 5. Percent developing invasive corpus and uterus NOS cancers before a specific age (Z), given cancer free at current age (Y), standard SEER areas, 1990–92.

Current Age (Y)	Develop Cancer by Age (Z)										
	10	20	30	40	50	60	70	80	90	Eventaully	1 in:
0	0.00	0.00	0.00	0.04	0.19	0.64	1.37	2.06	2.31	2.34	43
10		0.00	0.00	0.04	0.19	0.64	1.38	2.08	2.34	2.37	42
20			0.00	0.04	0.19	0.65	1.38	2.09	2.34	2.38	42
30				0.04	0.19	0.66	1.42	2.14	2.40	2.44	41
40					0.17	0.68	1.51	2.30	2.59	2.63	38
50						0.60	1.59	2.53	2.87	2.91	34
60							1.11	2.17	2.56	2.60	38
70								1.27	1.73	1.78	56

derivation of the number of newly developed cancer cases, ${}_5a_x$, is independent of the cancer prevalence correction factor, $l_x/{}^0l_x$. The second is that the lifetime risk estimate is always smaller when it is corrected for a surgical procedure. In the context of the second result, two competing factors are at play. When an intervening procedure eliminates individual's from the risk pool for a given disease, the lifetime probability of the disease decreases. However, at any given age interval since the denominator of at-risk individuals is reduced when those having the procedure are removed, the estimate of risk (events/at-risk population) will increase.

These two results hold exactly under the assumption of a uniform probability of an event during an interval, or approximately under an exponential assumption. We choose to present it here using the uniform assumption, which allows us to more clearly develop these two points and to compare our results directly with those found in the past (e.g., Seidman, et al., 1978).

Although the prevalence correction factor $l_x/{}^0l_x$ in equations (10) and (17) is presented as an improvement over previous calculations, lifetime risk estimates are independent of this correction. However, the prevalence correction factor is still needed for age-conditional risk estimates. Under the uniform assumption, we compute the number of newly developed cancer cases, ${}_5a_x$, in the interval $[x, x+5]$, as the number of cancer free persons at age x, ${}^0l_x$, multiplied by the probability of developing the cancer in the interval, ${}^0_5g_x$, multiplied by the probability of not dying of other causes prior to developing this cancer, $1-(1/2)\,{}^0_5q^0_x$; that is,

$$ {}_5a_x = ({}^0l_x)({}^0_5g_x)(1-(1/2)\,{}^0_5q^0_x), \tag{20}$$

With the prevalence correction factor from equation (10), we can write equation (20) as

$$ {}_5a_x = (l_x)({}_5g_x)(1-(1/2)\,{}^0_5q^0_x), \tag{21}$$

Equation (21) is calculated based on the total living rather than on the cancer free population. Seidman, et al. (1978), incorrectly calculated the number in the hypothetical cohort who develop cancer as

$$ {}_5a_x = ({}^0l_x)({}_5g_x)(1-(1/2)\,{}^0_5q^0_x), $$

Because the a's correctly calculation in equation (21) are independent of the prevalence correction and ${}^0l_0$ is a fixed number, the cumulative probability of developing cancer from birth, derived using equation (7), is independent of the prevalence correction. However, the probability of developing cancer from age j ($j > 0$), derived using equation (8), is not independent of the prevalence correction because ${}^0l_j$ depends on the prevalence correction.

Adjusting for a surgical procedure which removes individuals from risk of developing a given cancer leads to lifetime risk estimates which are smaller than unadjusted lifetime risk estimates. In the example in Section 4, adjusting for the prevalence of hysterectomy reduced the lifetime risk for corpus and uterus NOS cancers from 2.68% to 2.34%. Again, this is demonstrated under the assumption that the probability of an event occurs in the interval which follows an uniform distribution. When correcting for a surgical procedure (sp), equations (20) and (21) become

$$\begin{aligned} {}_5a_x^{\mathrm{sp}} &= ({}^0l_x)({}^0_5g_x)(1 - (1/2)\,{}^0_5q_x^0 - (1/2)\,{}^0_5f_x + (1/3)\,{}^0_5q_x^0\,{}^0_5f_x) \\ &= (l_x)({}_5g_x)(1 - (1/2)\,{}^0_5q_x^0 - (1/2)\,{}^0_5f_x + (1/3)\,{}^0_5q_x^0\,{}^0_5f_x). \end{aligned} \tag{22}$$

(The derivation of equation (22) is given in Appendix B.) Thus,

$$ {}_5a_x - {}_5a_x^{\mathrm{sp}} = (l_x)({}_5g_x)({}^0_5f_x((1/2) - (1/3)\,{}^0_5q_x^0)).$$

Since ${}^0_5q_x^0$ is a probability, it will not exceed one. Therefore,

$$((1/2) - (1/3)\,{}^0_5q_x^0) > 0.$$

This implies that

$$ {}_5a_x > {}_5a_x^{\mathrm{sp}}.$$

Hence, the lifetime risk calculation through equation (7) is smaller when a correction factor is used for a surgical procedure.

In equation (12), the calculation of the cancer free incidence rate ${}^0r_{95+}$ depends on the quantity ω. Here ω must be less than one for ${}^0r_{95+}$ to be greater than zero; that is, we must have

$$\frac{l_{95+}}{{}^0l_{95+}} < \frac{m_{95+}}{r_{95+}}.$$

Because the mortality rate m_{95+} from other causes is almost always larger than the incidence rate r_{95+} in the final open age interval, this condition is generally satisfied.

The cancer prevalence correction factor used in the calculation is derived cross-sectionally from current incidence and mortality data. We could have used a prevalence correction factor derived by following cohorts from an external data source, such as the Connecticut Tumor Registry (Connelly et al., 1968). This type of prevalence correction may better reflect rapidly changing incidence and mortality trends in cancer data, and deserves further investigation. However, an internal prevalence correction factor is often the only choice we have because most population based cancer registries have not existed long enough to provide accurate prevalence estimates. Based on Connecticut Tumor Registry data, Feldman et al. (1986)

found that 27 years of all newly diagnosed cancer cases were required for the prevalence proportion to be within five percent of the prevalence proportion based on a 47 year period.

In Section 2.3 we utilized the exponential assumption to define the probability of death. Based on the likelihood that the probability of death in the age interval follows an uniform distribution, then

$${}_5q_x = \frac{{}_5m_x}{(1/5)\{1+5(1-{}_5k_x)\}{}_5m_x},$$

and

$${}^0_5q^0_x = \frac{{}^0_5m^0_x}{(1/5)\{1+5(1-{}_5k_x)\}{}^0_5m^0_x}.$$

In these equations ${}_5k_x$ is the expected fraction of the 5-year-lived interval. By assuming an approximate uniform distribution of time at death for the interval $[x,\ x+5]$, then ${}_5k_x = 1/2$, and q is approximately the same as when derived under the exponential assumption. However, this assumption is perhaps an oversimplification, particularly in the first interval $[0, 5]$ where a large portion of the deaths are likely to occur within the first year. Previous studies suggest that a more appropriate value for k is near 0.1 (Chiang, 1972; Elandt-Johnson and Johnson, 1990). Referring again to the example in Section 4, we replace the q values in the derivation of lifetime and age-conditional risk estimates with those under an uniform distribution and ${}_5k_x = 0.1$. While this influences the number who die in the first interval, it has only a negligable influence on the risk of developing cancer. In Table 4, the number who die of other causes from age 0 to 4 changes from 92,716 to 92,374 and the lifetime risk of developing cervical cancer changes from 0.023437571 to 0.02343838.

It is important that the derivation of lifetime and age-conditional risk estimates be well understood by the statistical community so that these statistics can be properly computed and interpreted for the public. It is also important that the public be aware that these risk estimates are based on current age-specific cancer incidence, mortality, and, if applicable, surgery rates in the population, and are limited in that many factors will change as a baby born today ages over its lifetime. In this sense lifetime risk is similar to life expectancy since it is a hypothetically constructed index derived from current vital statistics. Because shorter-term age-conditional risk estimates are less susceptible to changes in incidence, mortality, and surgery rates in the future, these estimates may better reflect the risk for a person alive and cancer free in the population today.

Thus shorter-term age-conditional risk estimates better reflect people's "true" risk experience. In addition, they are more relevant to those approaching ages when cancer rates rise rapidly. On the other hand, lifetime risk estimates combine in a measure the current cancer incidence, mortality, and, when applicable, surgery rates in each age interval of the population. This measure may be useful, when compared to lifetime risk estimates from previous time periods, as a cancer progress indicator.

Other statistics have similar, yet contrasting features to lifetime and age-conditional risks. For example, cumulative incidence rates, which are used to approximate cumulative risk, assume that persons remain at risk for the period of interest, and are not subject to competing risks of death from other causes. Like lifetime risk, cumulative risk assumes that the current risk estimates remain stable. However, cumulative risk, as opposed to lifetime risk, is a cancer progress measure which is free of the influence of mortality from other causes. Nevertheless, these two statistics often provide an informative contrast. At every age, U.S. Blacks have higher age specific incidence rates for prostate cancer than whites. Thus even though Blacks have a larger cumulative risk than Whites, Whites have a larger lifetime risk because fewer Blacks reach the ages where prostate cancer rates rise rapidly.

By presenting a detailed description of the methodology for deriving lifetime and age-conditional risk estimates of developing cancer, we have attempted to clarify the interpretation and limitations of these statistics. Because of the frequency in which these risk measures appear in the popular press and are utilized by public health and cancer researchers, actuaries, health activists, and so on, it is particularly important that the statistical community understand their derivation and use.

Appendix A Derivation of ${}^{0}r_{95+}$

Part I. The Base Model

By definition

$$ {}^{0}r_{95+} = C_{95+}/{}^{0}L_{95+} = (L_{95+}/{}^{0}L_{95+})r_{95+}. \tag{A1} $$

When ${}^{0}L_{95+}$ is not directly available from the data, we assume that the person-years at risk remaining in the entire 95 year old cohort, p_{95+}, divided by the person-years at risk remaining among the cancer free cohort, ${}^{0}p_{95+}$, is equal to the ratio $L_{95+}/{}^{0}L_{95+}$, that is

$$ p_{95+}/{}^{0}p_{95+} = L_{95+}/{}^{0}L_{95+} \tag{A2} $$

where

$$ p_{95+} = l_{95+}(1/m_{95+}). \tag{A3} $$

The term $(1/m_{95+})$ is the expected number of person years remaining for someone age 95, under the exponential assumption. Similarly,

$$ {}^{0}p_{95+} = {}^{0}l_{95+}(1/({}^{0}m^{0}_{95+} + {}^{0}r_{95+})). \tag{A4} $$

Substituting (A2), (A3), (A4), into (A1) gives

$$ \begin{aligned} {}^{0}r_{95+} &= (p_{95+}/{}^{0}p_{95+})r_{95+} \\ &= (l_{95+}/m_{95+})[({}^{0}m^{0}_{95+} + {}^{0}r_{95+})/{}^{0}l_{95+}]r_{95+}. \end{aligned} $$

Solving for ${}^{0}r_{95+}$ we get

$$ {}^{0}r_{95+} = (\omega/(1-\omega)^{0}m^{0}_{x} $$

where

$$ \omega = [l_{95+}/(m_{95+} \times {}^{0}l_{95+})]r_{95+}. $$

Part II. The Extended Model

By definition

$$ {}^0r_{95+} = C_{95+}/{}^0L_{95+} = (L_{95+}/{}^0L_{95+})r_{95+}, \tag{A5}$$

and

$$ {}^0h^0_{95+} = (L_{95+}/{}^0L_{95+})h^0_{95+}.$$

When ${}^0L_{95+}$ is not directly available from the data, we have to derive both ${}^0r_{95+}$ and ${}^0h^0_{95+}$. We assume

$$p_{95+}/{}^0p_{95+} = L_{95+}/{}^0L_{95+}, \tag{A6}$$

where

$$ {}^0p_{95+} = {}^0l_{95+}(1/({}^0m^0_{95+} + {}^0r_{95+} + {}^0h^0_{95+})).$$

Then, using (A6) and (A3) in (A5) gives

$$\begin{aligned} {}^0r_{95+} &= (p_{95+}/{}^0p_{95+})r_{95+} \\ &= (l_{95+}/m_{95+})[({}^0m^0_{95+} + {}^0r_{95+} + {}^0h^0_{95+})/{}^0l_{95+}]r^0_{95+} \end{aligned}$$

and similarly

$$ {}^0h^0_{95+} = (l_{95+}/m_{95+})[({}^0m^0_{95+} + {}^0r_{95+} + {}^0h^0_{95+})/{}^0l_{95+}]h^0_{95+}.$$

We then solve for ${}^0r_{95+}$ and ${}^0h^0_{95+}$ simultaneously to get

$$ {}^or_{95+} = [\omega/(1 - \omega - \Psi)]^0m^0_{95+},$$

and

$$ {}^oh^0_{95+} = [\Psi/(1 - \omega - \Psi)]^0m^0_{95+},$$

where

$$\omega = (l_{95+} \times r_{95+})/(m_{95+} \times {}^0l_{95+}), \quad \Psi = (l_{95+} \times h^0_{95+})/(m_{95+} \times {}^0l_{95+}).$$

Appendix B Derivation of ${}^0a^{\text{sp}}_{95+}$

The derivation of the number of newly developed cancers in the five-year age intervals, corrected for hysterectomy (${}_5a^{\text{sp}}_x$) is presented here. We proceed on the assumption that the probability of an event occurs in the interval which follows an uniform distribution.

Let a five-year time period represent a unit of time, [0, 1], and the probability density function be denoted as $f(\cdot)$. In addition, let

X = the potential time a person is diagnosed with the cancer,
Y = the potential time a person dies from causes other than the cancer , and
Z = the potential time a person has a surgical procedure performed which
removes them from risk of developing the cancer, for reasons other than the ca

Note that "potential time" refers to the hypothetical time an event occurs in the absence of competing risks (Elandt-Johnson and Johnson, (1980)). Then,

$$f(X) = \begin{cases} {}^{0}_{5}g_x \text{ for } X \in [0, 1/{}^{0}_{5}g_x] \\ 0 \qquad \text{otherwise} \end{cases},$$

$$f(Y) = \begin{cases} {}^{0}_{5}q^{0}_{x} \text{ for } Y \in [0, 1/{}^{0}_{5}q^{0}_{x}] \\ 0 \qquad \text{otherwise} \end{cases},$$

and

$$f(Z) = \begin{cases} {}^{0}_{5}f \text{ for } Z \in [0, 1/{}^{0}_{5}f] \\ 0 \qquad \text{otherwise} \end{cases}.$$

Hence, the probability that a person is diagnosed with the cancer during a time period before death, or receiving a surgical procedure which removes them from being at risk of developing the cancer, is

$$\begin{aligned} &\Pr(X < 1,\ X < Y,\ X < Z) \\ &\quad = \int_0^1 \int_X^{1/{}^{0}_{5}q^{0}_{x}} \int_X^{1/{}^{0}_{5}f_x} {}^{0}_{5}g_x \times {}^{0}_{5}q^{0}_{x} \times {}^{0}_{5}f_x \, dz\, dy\, dx \\ &\quad = {}^{0}_{5}g_x \times {}^{0}_{5}q^{0}_{x} \times {}^{0}_{5}f_x \int_0^1 \left(\int_X^{1/{}^{0}_{5}q^{0}_{x}} 1\, dy \right) \left(\int_X^{1/{}^{0}_{5}f_x} 1\, dz \right) dx \\ &\quad = {}^{0}_{5}g_x (1 - (1/2)({}^{0}_{5}q^{0}_{x} + {}^{0}_{5}f_x) + (1/3)\, {}^{0}_{5}q^{0}_{x} \times {}^{0}_{5}f_x) \\ &\quad = {}^{0}_{5}g_x (1 - (1/2)\, {}^{0}_{5}q^{0}_{x} - (1/2)\, {}^{0}_{5}f_x + (1/3)\, {}^{0}_{5}q_x \times {}^{0}_{5}f_x). \end{aligned}$$

References

A. P. Bender, J. Punyko, A. N. Williams, and S. A. Bushhouse, "A standard person-years approach to estimating lifetime cancer risk. The section of chronic disease and environmental epidemiology, Minnesota Department of Health," *Cancer Causes Control* vol. 3 pp. 69–75, 1992.

S. Blakeslee, "Faulty math heightens fears of breast cancer," *The New York Times*, March 15, 1992, Section 4, pp. 1,6.

B. W. Brown, C. Brauner, and M. C. Minnote, "Noncancer deaths in white adult cancer patients," *J Natl Cancer Inst* vol. 85 pp. 979–987, 1993.

C. C. Chiang, "On constructing current life tables," *J Amer Statist Assoc* vol. 67 pp. 538–541, 1972.

R. R. Connelly, P. C. Campbell, and H. Eisenberg "Central registry of cancer cases in Connecticut," *Public Health Rep.* vol. 83 pp. 386–390, 1968.

R. C. Elandt-Johnson and N. L. Johnson, *Survival Models and Data Analysis*, John Wiley and Sons: New York, 1980, p. 271.

A. R. Feldman, L. Kessler, M. H. Myers, and M. D. Naughton, "The prevalence of cancer: Estimates based on the Connecticut tumor registry," *N Engl J Med* vol. 315 pp. 1394–1397, 1986.

E. J. Feuer, L. M. Wun, and C. C. Boring, "Probability of developing cancer," in B. A. Miller, L. A. G. Ries, B. R. Hankey et al. (eds). *Cancer Statistics Review: 1973-1989*, National Cancer Institute, NIH Pub. No 92-2789, 1992, XXX.1-8.

E. J. Feuer, L. M. Wun, C. C. Boring, et al., "The lifetime risk of developing breast cancer," *J Natl Cancer Inst* vol. 85 pp. 892–897, 1993.

E. J. Feuer and L. M. Wun, "DEVCAN: Probability of DEVeloping CANcer software," version 3.11, National Cancer Institute, 1994.

M.H. Gail, "A review and critique of some models used in competing risk analysis," *Biometrics* vol. 31 pp. 209–222, 1975.

M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer and J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *J Natl Cancer Inst* vol. 81 pp. 1879–1886, 1989.

I. D. Goldberg, M. L. Levin, P. R. Gerhardt, V. H. Handy and R. E. Cashman, "The probability of developing cancer," *J Natl Cancer Inst* vol. 17 pp. 155–173, 1956.

E. J. Graves, "National hospital discharge survey: Annual summary. 1990," DHHS publication no. (PHS) 92-1773. (Vital and health statistics 13, no. 112). Hyattsville, Maryland: National Center for Health Statistics, 1992.

M. Kramer, M. Von Korff and L. Kessler, "The lifetime prevalence of mental disorders: Estimation, uses and limitations," *Psychol Med* vol. 10 pp. 429–435, 1980.

J. L. Lyon and J. W. Gardner, "The rising frequency of hysterectomy: Its effect on uterine cancer rates," *Am J Epidemiol* vol. 105 pp. 439–443, 1977.

R. Pokras, "Hysterectomy, past, present, and future," *Stat Bull Metrop Insur Co* vol. 70 pp. 12–21, 1989.

R. M. Merrill and E. J. Feuer, "Risk-adjusted cancer incidence rates," *Cancer Causes Control* vol. 7 pp. 553–561, 1996.

One in nine. *Washington Post*, March 17, 1992, Section A, p. 16.

L. A. G. Ries, C. L. Kosary, B. F. Hankey, B. A. Miller, A. Harras, and B. K. Edwards (eds.), *SEER Cancer Statistics Review, 1973-1994: Tables and Graphs*, NIH Pub. No. 97-2789. Bethesda, MD: National Cancer Institute, 1997.

L. Schwab, "Breast cancer challenge sounded at Capitol," *J Natl Cancer Inst* vol. 83 p. 914, 1991.

H. Seidman, E. Silverberg and A. Bodden, "Probabilities of eventually developing and dying of cancer (Risk among persons previously undiagnosed with cancer)," *CA Cancer J Clin* vol. 28 pp. 33–46, 1978.

S. L. Parker, T. Tong, S. Bolden and P. A. Wingo, "Cancer statistics, 1997," *CA Cancer J Clin* vol. 47 pp. 5–27, 1997.

M. S. Zdeb, "The probability of developing cancer," *Am J Epidemiol* vol. 106 pp. 6–16, 1977.