Human and Ecological Risk Assessment: Vol. 7, No. 6, pp. 000-000 (2001)

Age Distribution of Cancer: The Incidence Turnover at Old Age

Francesco Pompei[+] and Richard Wilson*
Department of Physics, Harvard University, Jefferson Laboratories, Cambridge, MA 02138; *Tel (voice): 617-495-3387, Tel(fax): 617-495-0416; [+] Tel (voice): 617-923-9900 x201, Tel(fax): 617-923-9911; Email: * wilson@huhepl.harvard.edu, [+]fpompei@post.harvard.edu

ABSTRACT

In recent years data on cancer incidence in the USA, the Netherlands, and in Hong Kong indicate a flattening and perhaps a turnover at advanced age, but no model has been successful in fitting this data and thus providing clues to the underlying biology. In this work we assume these data are reliable and free from bias. We find that a Beta distribution fits SEER age-specific cancer incidence data for all adult cancers extremely well, and its interpretation as a model leads to the possibility that there is a beneficial cancer extinction process that becomes important at elevated age. Particularly evident from the data is the apparent remarkable uniformity of adult cancers peaking in incidence at about the same age, including cancers in other countries. Possible biological mechanisms include increasing apoptosis and cell senescence with age. Further, the model suggests that cancer is not inevitable at advanced age, but reaches a maximum cumulative probability of affliction with any cancer of about 70% for men and 53% for women in the US, and much smaller values for individual cancers.

Key Words: cancer incidence, age, cancer model, multistage, clonal expansion.

INTRODUCTION

It is well known that most cancers arise late in life. There is also substantial evidence that there is a latency period between the time of the initiation of a cancer to its observation. An early model was that the cancer cells multiply exponentially, at a slow but steady rate, and that only when they reach a certain critical number can the cancer be identified. The latency is then the time for this multiplication to occur. The assumption that cancers may be initiated throughout life leads naturally to observed age specific cancer incidence $I(t)$ increasing exponentially with age t as

[+] Corresponding author

1080-7039/01/$.50

$I(t)=Ae^{bt}$. The study of the age distribution of cancers began with national mortality data records of the deaths caused by cancers. It soon became apparent that the cancer death rate increases less steeply with age than the exponential.

Nordling (1953) and Armitage and Doll (1954), working with national mortality data in the UK, proposed an alternative multistage theory of cancer to describe the age distribution data. According to this theory, the cell multiplication is assumed to be rapid and the time from initiation to cancer observation (the latent period) is assumed to be the time of passing several discrete stages. This leads to the formula $I(t)=at^{k-1}$ or $ln\ I = ln\ a + (k\text{-}1)ln\ t$, where k is the number of stages and a includes various factors representing environmental exposure, genetic susceptibility, and dietary factors. Armitage and Doll successfully fitted age specific cancer mortality rates, which they assumed to approximately represent age specific incidence rates, from several sites and countries and found fits with values of k between 4 and 8. They omitted death rates above age 75, arguing that at such an advanced age, physicians would tend to assign the nebulous "old age" as the cause of death rather than make a more careful diagnosis. They therefore ignored the apparent flattening in age specific mortality rates in the data.

In the years since 1947, cancer registries recording incidence data have much improved. Also, the increase in survival times and cure rates for many cancers has made it desirable to examine incidence rates rather than death rates as indicators of biological mechanisms. Great improvements in the collection of incidence data suggest to us that the concerns of Armitage and Doll about using data from ages over 75, and using incidence data at all, may now be resolved. The reader must be warned that the conclusions of this paper depend critically on the assumption of validity of the modern incidence data, which is discussed further later in this paper.

This study follows the same steps as the cancer modelers of the 1950s: (1) taking the most recent data including the turnover to be considered reliable (SEER data, Reis *et al.* 2000); (2) attempting to fit them with as simple a model as necessary to obtain a good fit; (3) comparing curve shapes to other cancers and data from other countries; then (4) discussing possible clues to the underlying biology implied by the model. It is found that a good fit of all adult cancers can be made with a form of Beta distribution (Olkin *et al.* 1978) for age-specific incidence: $I(t)=(\alpha t)^{k-1}(1\text{-}\beta t)$, which includes 3 arbitrary constants. We extrapolate this distribution to older ages (~100 years) where few data exist and explore the implications of assuming the reliability of the model implied by the fits. As shown in Figure 1, a Beta fit to the SEER data for all cancers is a very different fit than the curves for the two historically important models, the A-D power law and the MVK clonal expansion models (both to be discussed in more detail) which have been frequently analyzed for insights to biological causes of cancer. Since there is a large difference, the Beta fit might be evaluated as a model, and its biological implications are briefly explored.

METHODS

SEER Data

This study takes the most recent age specific incidence data (1993 to 97) from the Surveillance, Epidemiology, and End Results (SEER) Program, based within the Cancer Surveillance Research Program at the National Cancer Institute (Ries *et al.*

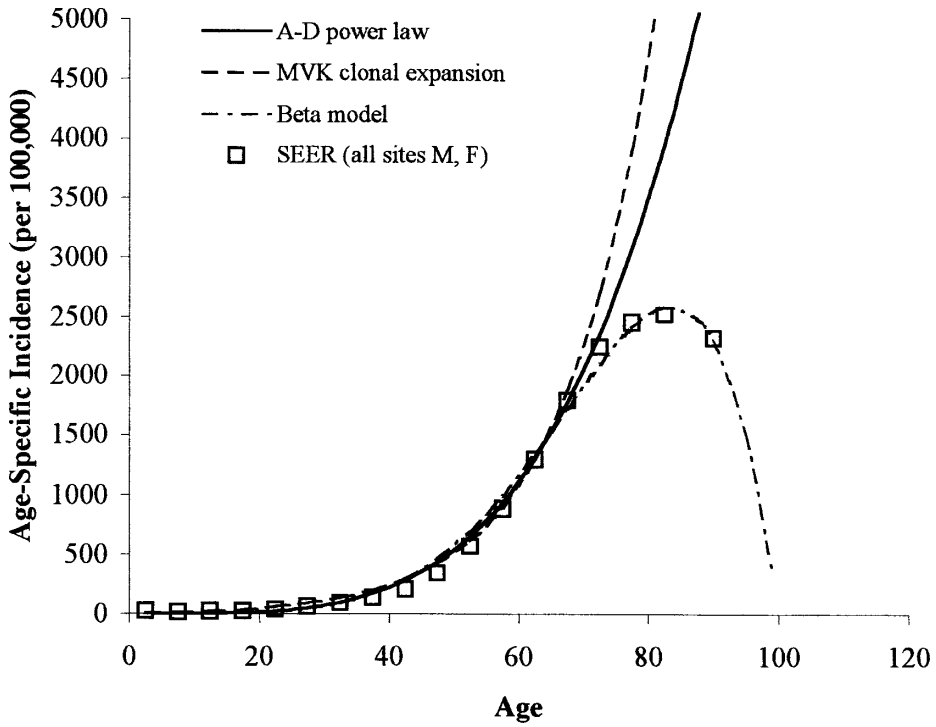

Figure 1. **Age specific incidence vs. age curve shapes for the two major historical model types, compared to the Beta model and SEER data for combined male and female cancers.**

2000). Established by the National Cancer Act of 1971, the SEER program routinely collects cancer incidence and mortality data from designated population based cancer registries in various areas of the country, representing about 14% of the US population. The cancer site and histology are coded according to the International Classification of Diseases for Oncology, second edition (ICD-O-2) (Percy *et al.* 1990).

We emphasize that the reliability of the SEER data is central to this work, and the conclusions depend on an explicit acceptance of the data as an accurate representation of the actual incidence in the US population. SEER follow a number of careful procedures to insure the quality of the data, including "abstracting records for resident cancer patients seen in every hospital both inside and outside each coverage area; searches of records of private laboratories, radiotherapy units, nursing homes, and other health services units that provide diagnostic service to ensure complete ascertainment of cases; records data on all newly diagnosed cancers, including selected patient demographics, primary site, morphology, diagnostic confirmation, extent of disease, and first course of cancer-directed therapy; and conduct periodic quality control studies to correct errors."

Age-specific cancer incidence is defined by SEER as $Cancer\ incidence=(d_i/n_i)*100,000$; where i = the 18 age groups 0-4, 5-9, …, 85+; d_i = number of new cancers diagnosed in age group i; n_i = person-years in group i. The denominator used by SEER represents the entire population in the relevant age group, *including all who have been*

diagnosed with the cancer at an earlier age and have not yet died of that cancer. This point will be important in discussing age-specific incidence data when interpreted as a hazard function (see Appendix).

The age-specific incidence data is organized as entries in 5-year age intervals starting from 0 to 4 to 80 to 84, ending with an 85+ category. For all intervals except the last, the center age was considered as representing that interval. For the 85+ category, a weighted mean value of 90 was computed from life tables (National Vital Statistics Report 1999) for persons living beyond 85, and used for the 85+ category. The database for age-specific cancer incidence includes data for 19 male and 21 female primary cancer sites in addition to all sites combined. No attempt was made to correct the data for population cohort effects such as hysterectomy, which would remove people from the denominator; or smoking status, which would provide variable sensitivities.

Comparisons to Other Datasets

Three other data sets were examined to assess the validity of the model at higher ages than reported by SEER, and to people from markedly different gene pools, diet, and environment. A study by de Rijke *et al.* (2000) presenting cancer incidence data to age 95+ for the Dutch population over the period 1989 to 1995 includes data at higher ages than SEER and from a different culture and environment. The cancer registries that the authors rely upon have been confirmed to have high (96.2%) completeness even in the highest age groups, and are considered reliable. For each age interval reported by de Rijke, the weighted mean age for that group was computed and used for all analyses and figures. Since the absolute numbers of cancer incidences are much smaller than the SEER data, particularly in the higher age groups, error bars representing ±2 SEM are indicated in the figures.

A cancer incidence dataset (Parkin *et al.* 1997) for the Hong Kong population over the period 1988 to 1992 provides 35,000 cancer incidences over 50 organ sites. The population, 98% of whom are Chinese, and 90% from a single nearby province in China, provides a comparison for a different gene pool, culture, and environment than Americans or Europeans. Participation in the cancer registry is voluntary, but data collection processes and checking procedures are believed to be effective in ensuring reliable data. Parkin *et al.*'s analysis indicates that its method of site incidence tabulation results in incidences within 5% of that employing the SEER method. The data extend over the same age range as SEER, ending at 85+. Only data for six major sites are examined.

A study by Saltzstein *et al.* (1998) examined cancer incidence recorded for the 35 million people of California over the period 1988 to 1993, and reported age-specific cancer rates in 5-year age groups from 50 to 54 to 95 to 99 and ≥100 years old. This study included 14,086 cancer patients over age 90, 70.8% of which had histological confirmation of their cancer diagnoses, compared to 94.5% of those less than 90. Although cancer cases in California are a major component of the SEER data and thus are expected to be similar, the investigators specifically designed their study to examine the incidences for older age groups than SEER report.

Beta Function Selection for Fits

Historically, a good fit to incidence data up to age 60 or 70 is given by $I(t)=\lambda t^{k-1}$. This suggests a modification of this formula to include a factor to produce the turnover. One possibility is a form of the Beta probability density function, described in statistics texts as: $f(x)=\lambda x^{k-1}(1-x);\ 0\leq x\leq 1$. We parameterize the Beta function by $x=\beta t$, giving: $I(t)=(\alpha t)^{k-1}(1-\beta t)*100{,}000;\ 0\leq t\leq \beta-1$, and we find a good fit for the SEER data by adjusting the constants α, k-1, and β. Although an additional arbitrary constant always enables a better fit, we suggest that the additional factor producing the turnover might represent a cancer extinction process. Unlike the mathematical models (described later) that have been used in the past to fit the incidence data, the Beta function has value 0 for $t\geq\beta-1$, thus suggesting the possibility of a limit to the cumulative probability of cancer that is less than one. The full derivation is in the Appendix.

The SEER cancer sites are divided into the 17 non-gender-specific sites (unrelated to reproductive organs), and the 6 gender-specific sites (related to reproductive organs) to be separately analyzed. The Beta fit to the 17 non-gender sites are performed with $t = age$. As first suggested by Armitage and Doll (1954) the sex organs may have different timing of carcinogenic influences compared to the non-gender-specific sites due to sexual maturity and activity. The simplest assumption is that the carcinogenic influences start at sexual maturity, taken as age 15. Thus for the 6 gender-specific sites the fits are performed with $t = (age-15) \geq 0$.

Goodness of Fit

We employ a fraction of variance method introduced to cancer modeling by Cox (1995). The fits were produced by manipulation of the three variables of the Beta model α, $k-1$, β to maximize the "Fit" value, which we define as the fraction of the variance in the observed data points accounted for by the model. The expression employed is $Fit = 1 - \{E[(\underline{O}-\underline{M})^2]/ E[(\underline{O}-\mu)^2]\} = 1- \{[\Sigma(o_i-m_i)^2]/ [\Sigma(o_i-\mu)^2]\}$, each summation taken over $i= 1$ to r, where $\underline{O}$ and $\underline{M}$ are the observations and model results random variables respectively, μ is the mean of all of the age-group observations for that cancer, o_i and m_i the observed and modeled values for each age group for that cancer, and r the number of age group data points to be fit. As indicated by the equation, a perfect model ($\Sigma(o_i-m_i)= 0$) gives a *Fit* value of unity, since the model fully accounts for all variance of the observations from the mean, and no model at all gives a value of zero, since it accounts for none of the variance.

No attempt was made to include the effect of the uncertainty in the mean for each data point value, since the SEER data is a record of about 35 million people (14% of the U.S. population), about 1% of which is in the smallest (85+) age group. Thus the maximum sampling error is about 1 per 100,000, which is smaller than the variability of the data for individual cancers by one to three orders of magnitude. The Beta function curve fit is extended a few years beyond the end of the SEER data in order to clearly show the location of the predicted peak in incidence. It should be emphasized that the SEER data by themselves do not show a peak for all sites within the age range reported, but when data are available to age ~100 (as in the Dutch and California data) a peak occurs for all organ sites reported.

A possible source of error in modeling the SEER data are birth cohort effects, in which persons in certain age groups are exposed to a non time-homogeneous cause or new diagnosis of cancer, such as popularity of smoking, introduction of a new diagnostic test, or a cataclysmic singularity in exposure to a carcinogen such as Hiroshima and Nagasaki. We have not attempted to correct any of the data for these effects, and although they might be important in modeling individual cancers, the main conclusions of this work are based on all 35 of the adult cancers, which are unlikely to be uniformly distorted by birth cohort effects.

RESULTS

Fits of the Beta Function to SEER Data

Figures 2(a-q) present the SEER data and Beta fits for each of the 17 non-gender-specific cancer sites for both males and females. While 31 of the 34 fits can be seen to be quite good (*Fit* values near 1), male and female Hodgkin's disease, and female thyroid cancer appear to be significantly different cancer types than those which are central to this work. The

Fit values for the 31 cancers range from 0.93 to 1.00 with a mean of 0.97 of the variance accounted for by the Beta function fit. For comparison, the A-D power law model of figure 1 produces modeled fraction of variance fit values of 0.99, 0.94, 0.69, and –0.3 for ages 0 to 74, 0 to 79, 0 to 84, and 0 to 90, respectively, for male liver cancer (a 1% cumulative incidence cancer, where the A-D approximation is mathematically accurate, as discussed below). The corresponding values for the Beta fit are 1.00, 1.00, 0.99, and 0.99 for the same cancer over the same age ranges.

Figure 2(r) shows the total incidence for all 17 non-gender-specific sites for males and females separately, created by summing the SEER data and the Beta fits for each age category (not a true probability but a commonly used approximation). Clearly the male and female incidence curves have the same shape, both reaching a peak at about age 80, but a factor of two difference in incidence. Figures 3(a-f) present the Beta fits for the six gender-specific sites, which are all based on *t=0* at age 15. All four of the female site fits are quite good (mean *Fit* value = 0.97), but the two male sites are not quite as good (mean *Fit* value = 0.92). Testicular cancer (*Fit* = 0.87) clearly is a very different cancer. Prostate cancer (*Fit* = 0.96) might be somewhat influenced by the SEER data itself, which the SEER investigators warn is heavily influenced by the prostate specific antigen (PSA) test entering into common use over the last few years. The SEER reported overall age-adjusted incidence rate shows a distinct "bump" in the years 1989 to 1996, but the age distribution of this bump is not reported.

Tables 1 and 2 present the tabulation of the Beta parameters for the fits for males and females respectively ranked by peak incidence, and calculated implications compared to the SEER data. The α parameter varies with the ranking of peak incidence when *(k-1)* values are similar, departing somewhat from this ranking when *(k-1)* is different. The β parameter is remarkably consistent, varying by only about 20% for the 35 adult cancers, even as the peak incidences vary by a factor of 100. Also, the value of α is always less than the value of β, suggesting that the probabilities of the *(k-1)* uniform random variables representing cancer creation are always less

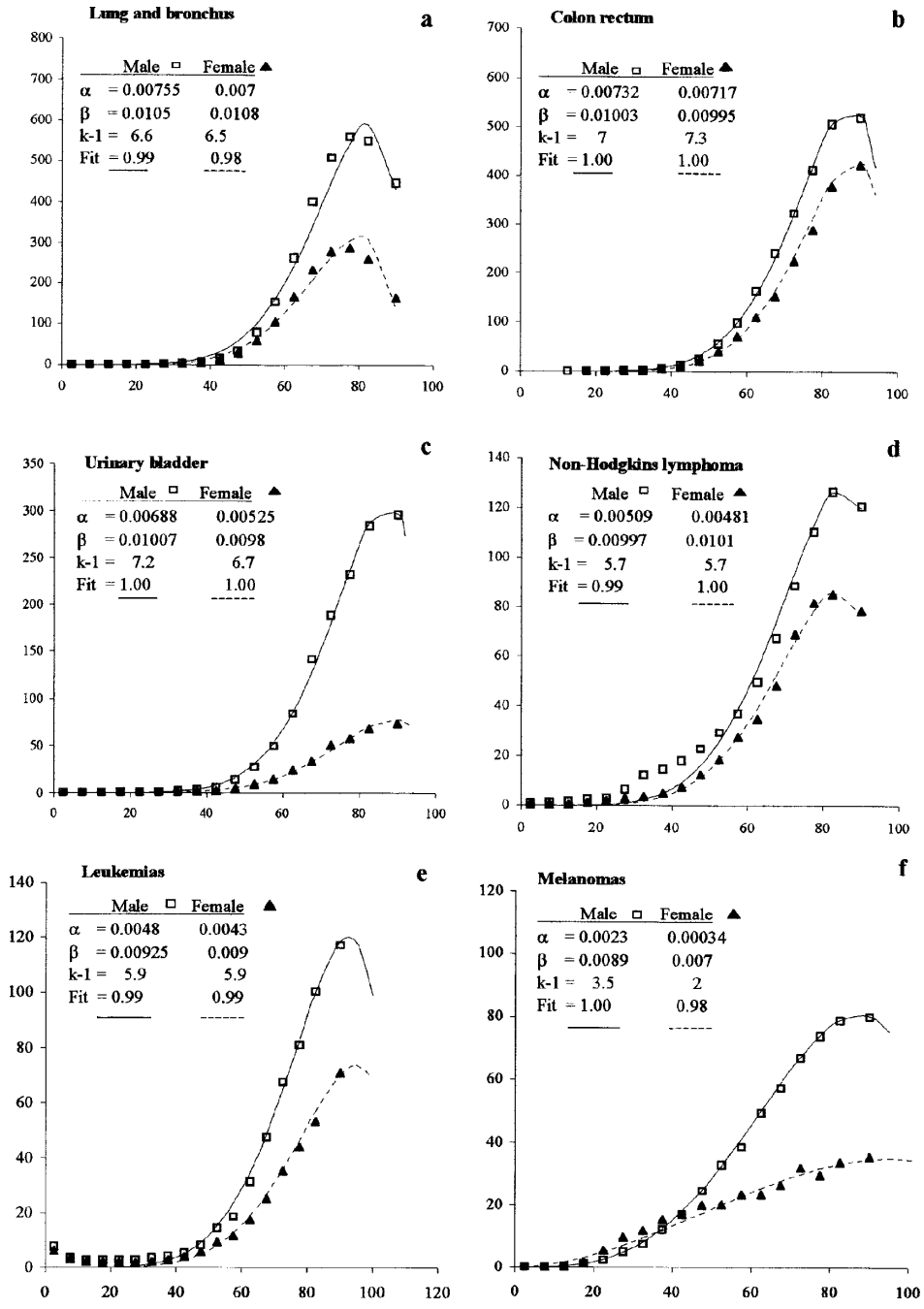

Figure 2a-r. **Age specific incidence (per 100,000) vs. age for males and females.** Beta distribution fits of SEER (Reis *et al.* 2000) data for non-gender-specific sites. Parameter values are listed for the Beta function form: $I(t) = (\alpha t)^{k-1}(1-\beta t)*100{,}000$. The fit values are calculated as the fraction of the variance of the observed data, which are accounted for by the Beta model with the listed parameter values.

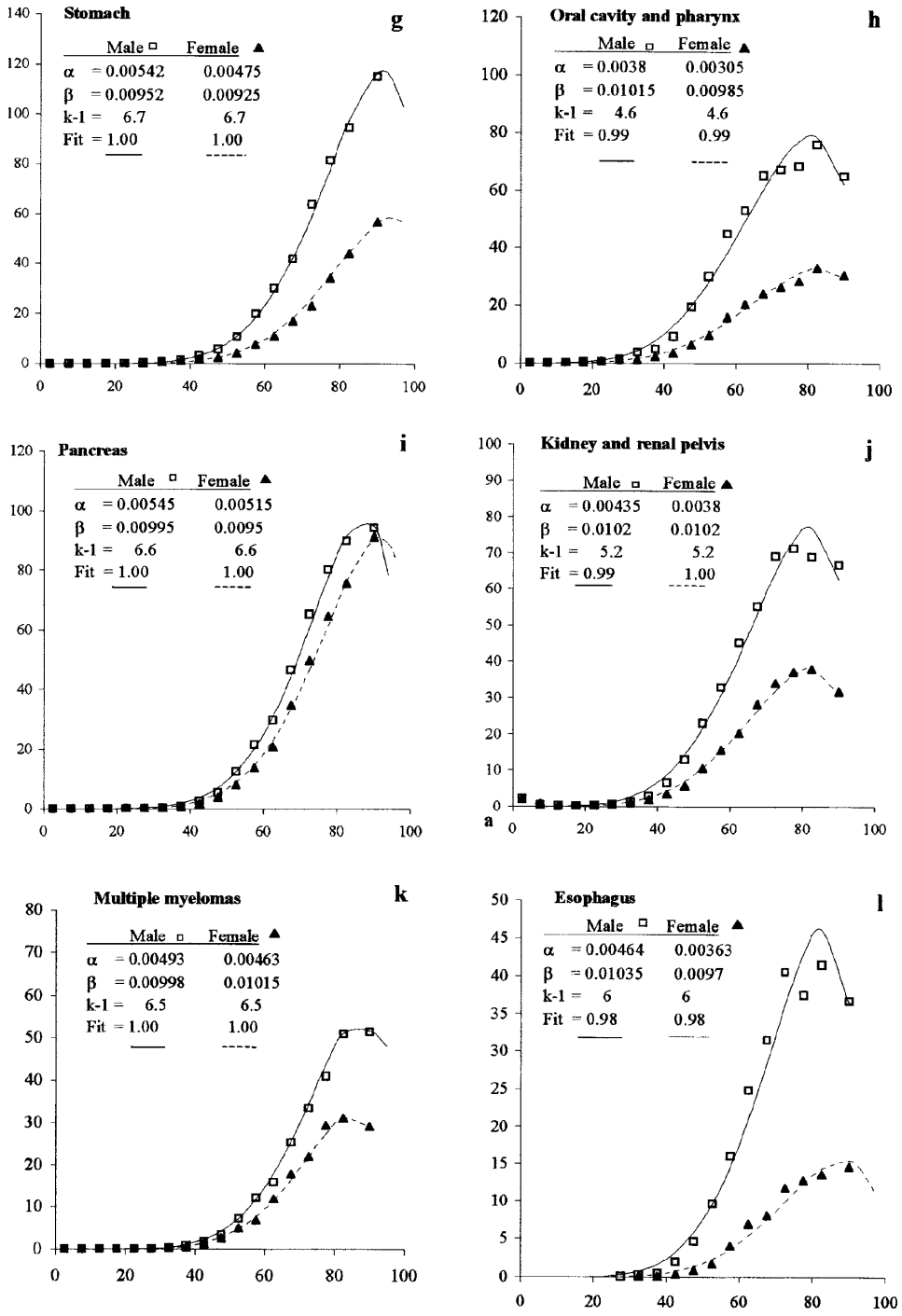

FIGURE 2. (continued)

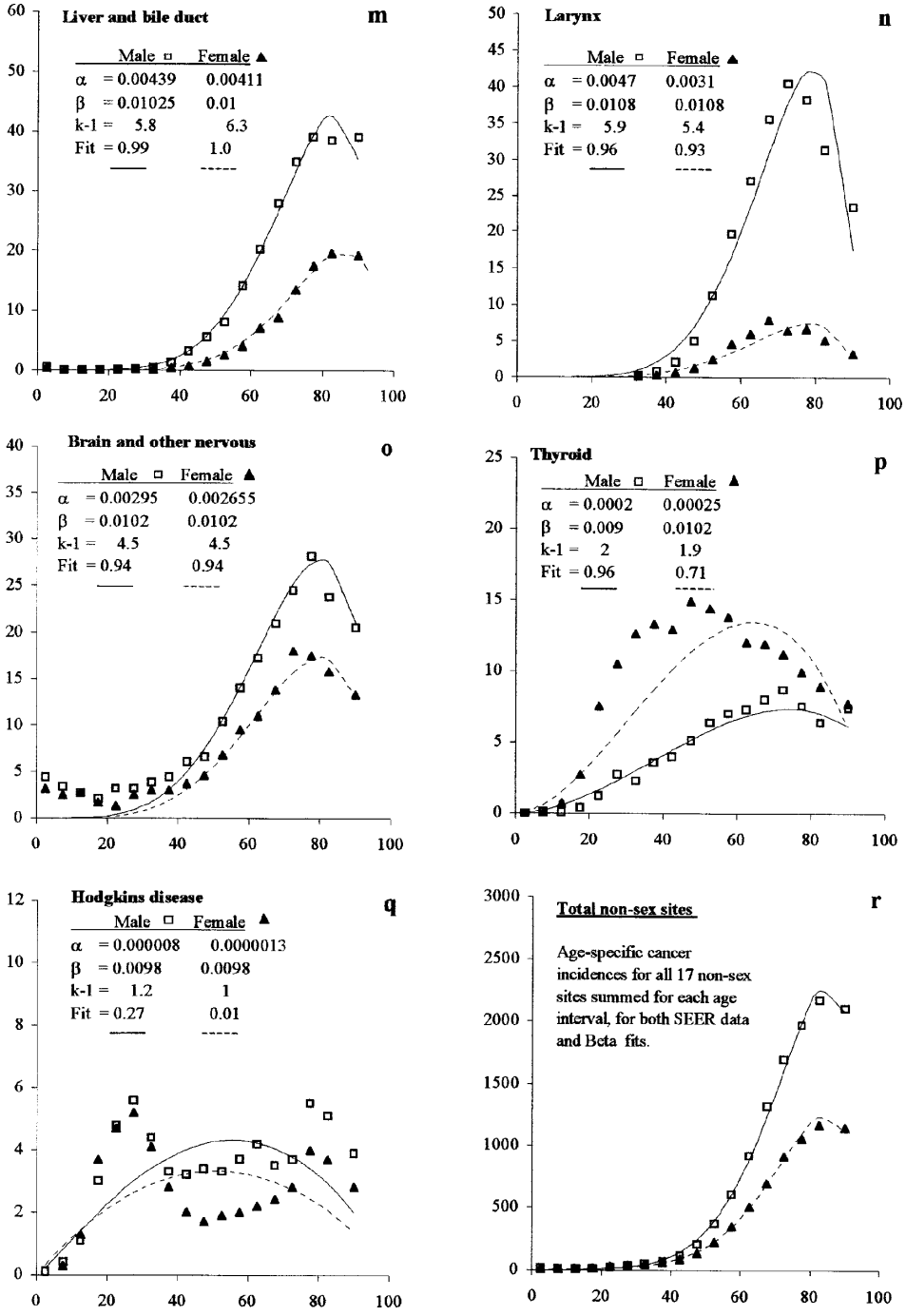

FIGURE 2. (continued)

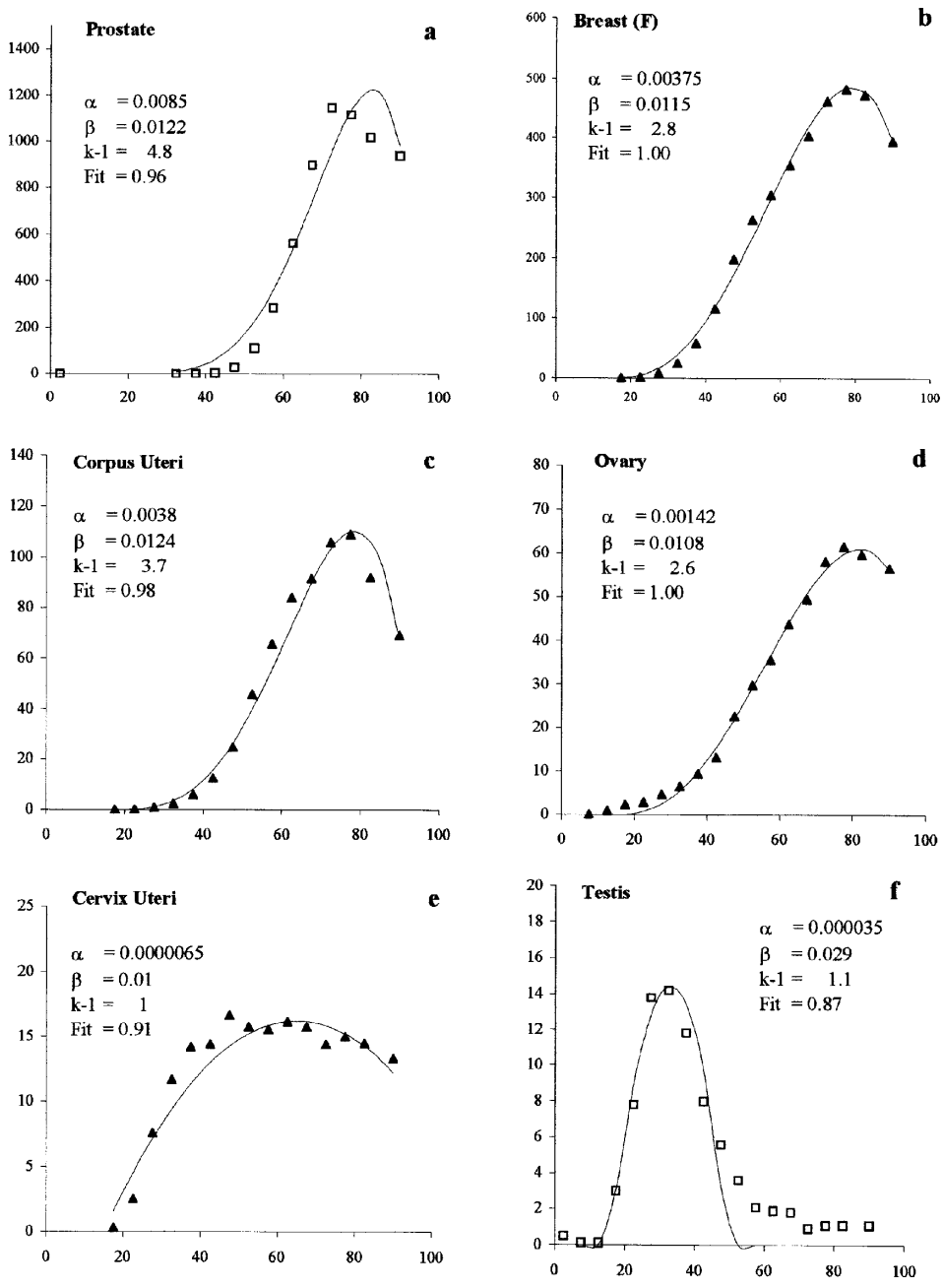

Figure 3a-f. Age specific incidence (per 100,000) vs. age. Beta distribution fits of SEER (Reis *et al.* 2000) data for gender-specific sites. Parameter values are listed for the Beta function form: $I(t) = (\alpha t)^{k-1}(1-\beta t)*100{,}000$, where $t = age-15$. The fit values are calculated as the fraction of the variance of the observed data, which are accounted for by the Beta model with the listed parameter values.

Table 1. Beta fits to SEER data for males: Parameter values and their implications compared to SEER data.

Site	α ($\times 10^2$)	k-1	β ($\times 10^2$)	Fit (fraction of variance modeled)	Peak age-specific incidence (per 100k) $I(t_p)$; [SEER]	Age at peak incidence t_p; [SEER]	Age at zero incidence $\beta^{1} = t_o$	Cumulative probability* over lifespan P_{cl}; [SEER]$^{++}$
Non-gender-specific sites								
Lung and bronchus	0.755	6.6	1.05	0.986	588; [560]	83; [78]	95	0.165; [0.153]$^{++}$
Colon and rectum	0.732	7	1.003	1.000	541; [519]$^{+}$	87; [90]$^{+}$	100	0.153; [0.132]$^{++}$
Urinary bladder	0.688	7.2	1.007	0.998	308; [296]$^{+}$	87; [90]$^{+}$	99	0.085; [0.074]$^{++}$
Non-Hodgkin's lymphoma	0.509	5.7	0.997	0.985	129; [127]	85; [83]	100	0.042; [0.039]$^{++}$
Leukemias	0.48	5.9	0.925	0.994	120; [117]$^{+}$	92; [90]$^{+}$	108	0.041; [0.029]$^{++}$
Stomach	0.542	6.7	0.952	0.998	117; [117]$^{+}$	91; [90]$^{+}$	105	0.036; [0.026]$^{++}$
Pancreas	0.545	6.6	0.995	0.998	98; [94]$^{+}$	87; [90]$^{+}$	101	0.029; [0.025]$^{++}$
Melanomas of skin	0.23	3.5	0.89	0.999	81; [80]$^{+}$	87; [90]$^{+}$	112	0.040; [0.029]$^{++}$
Oral Cavity and Pharynx	0.38	4.6	1.015	0.986	79; [76]	81; [83]	99	0.029; [0.027]$^{++}$
Kidney and renal pelvis	0.435	5.2	1.02	0.990	77; [71]	82; [78]	98	0.026; [0.025]$^{++}$
Multiple myeloma	0.493	6.5	0.998	0.998	54; [51]	87; [85]	100	0.016; [0.013]$^{++}$
Esophagus	0.464	6	1.035	0.981	46; [42]	83; [83]	97	0.014; [0.013]$^{++}$
Liver and bile duct	0.439	5.8	1.025	0.991	43; [39]	83; [83]	98	0.013; [0.013]$^{++}$
Larynx	0.47	5.9	1.08	0.960	42; [41]	79; [73]	93	0.013; [0.011]$^{++}$
Brain and other nervous	0.295	4.5	1.02	0.942	28; [28]	80; [78]	98	0.010; [0.011]$^{++}$
Thyroid	0.02	2	0.9	0.956	7; [9]	74; [73]	111	0.005; [0.004]$^{++}$
Hodgkin's disease	0.0008	1.2	0.98	0.267	4; [6]	56; [28]	102	0.003; [0.003]$^{++}$
Gender-specific sites								
Prostate	0.9500	5.8	1.250	0.6466	1227; [1149]	83; [73]	97	0.367; [0.329]$^{++}$
Testis	0.0035	1.1	2.900	0.7289	14; [14]	33; [33]	49	0.003; [0.004]$^{++}$
At least one cancer $\sum_{all} I(t_p) =$					3603; [3436]	$1-\prod_{all}(1-P_c) =$		0.704; [0.652]$^{++}$
At least one non-gender-specific cancer $\sum_{non-sex} I(t_p) =$					2362; [2273]	$1-\prod_{non-sex}(1-P_c) =$		0.531; [0.480]$^{++}$

* $P_c = \int_0^{\beta^{1}} (\alpha t)^{k-1}(1-\beta t)\,dt$

$^{+}$ Indicates SEER data that does not record a peak prior to the highest age category, which is mean 90 years.
$^{++}$ Indicates integration of SEER data up to age 90.

Table 2. Beta fits to SEER data for females: Parameter values and their implications compared to SEER data.

Site	α (x10^2)	k-1	β (x10^2)	Goodness of Fit	Peak age-specific incidence (per 100k) $I(t_p)$; [SEER]	Age at peak incidence t_p; [SEER]	Age at zero incidence $\beta^{-1} = t_o$	Cumulative probability* over lifespan P_c; [SEER]$^{++}$
Non-gender-specific sites								
Colon and rectum	0.717	7.3	0.995	0.997	432; [423]$^{+}$	88; [90]$^{+}$	101	0.119; [0.097]$^{++}$
Lung and bronchus	0.7	6.5	1.08	0.978	314; [287]	80; [77.5]	93	0.087; [0.084]$^{++}$
Pancreas	0.515	6.6	0.95	0.998	91; [91]$^{+}$	91; [90]$^{+}$	105	0.028; [0.021]$^{++}$
Non-Hodgkin's lymphoma	0.481	5.7	1.01	0.996	87; [85]	84; [82.5]	99	0.028; [0.026]$^{++}$
Urinary bladder	0.525	6.7	0.98	0.995	78; [73]$^{+}$	89; [90]$^{+}$	102	0.023; [0.019]$^{++}$
Leukemias	0.43	5.9	0.9	0.987	74; [71]$^{+}$	95; [90]$^{+}$	111	0.026; [0.017]$^{++}$
Stomach	0.475	6.7	0.925	0.997	59; [57]$^{+}$	94; [90]$^{+}$	108	0.019; [0.012]$^{++}$
Kidney and renal pelvis	0.38	5.2	1.02	0.995	38; [38]	82; [82.5]	98	0.013; [0.012]$^{++}$
Melanomas of skin	0.034	2	0.7	0.977	35; [36]$^{+}$	95; [90]$^{+}$	143	0.028; [0.016]$^{++}$
Oral Cavity and Pharynx	0.305	4.6	0.985	0.993	33; [33]	83; [82.5]	102	0.012; [0.011]$^{++}$
Multiple myeloma	0.463	6.5	1.015	0.997	32; [31]	85; [82.5]	99	0.009; [0.009]$^{++}$
Liver and bile duct	0.411	6.3	1	0.996	20; [20]	86; [82.5]	100	0.006; [0.005]$^{++}$
Brain and other nervous	0.2655	4.5	1.02	0.943	17; [18]	80; [77.5]	98	0.006; [0.007]$^{++}$
Esophagus	0.363	6	0.97	0.985	16; [15]$^{+}$	88; [90]$^{+}$	103	0.005; [0.004]$^{++}$
Thyroid	0.025	1.9	1.02	0.706	13; [15]	64; [47.5]	98	0.008; [0.008]$^{++}$
Larynx	0.31	5.4	1.08	0.928	7; [8]	78; [67.5]	93	0.002; [0.002]$^{++}$
Hodgkin's disease	0.0001	1	0.98	0.008	3; [5]	51; [22.5]	102	0.002; [0.002]$^{++}$
Gender-specific sites								
Breast	0.375	2.8	1.15	0.995	486; [482]	79; [77.5]	102	0.207; [0.187]$^{++}$
Corpus uteri	0.38	3.7	1.24	0.980	110; [109]	79; [77.5]	96	0.038; [0.033]$^{++}$
Ovary	0.142	2.6	1.08	0.996	61; [62]	82; [77.5]	108	0.029; [0.024]$^{++}$
Cervix uteri	0.0007	1	1	0.911	16; [17]	65; [47.5]	115	0.011; [0.010]$^{++}$
At least one cancer $\sum_{all} I(t_p) =$					2022; [1976]	$1-\prod_{all}(1-P_c) =$		0.526; [0.475]$^{++}$
At least one non-gender-specific cancer $\sum_{non-sex} I(t_p) =$					1349; [1306]	$1-\prod_{non-sex}(1-P_c) =$		0.354; [0.305]$^{++}$

$$^{*}P_c = \int_0^{\beta^{-1}} (\alpha t)^{k-1}(1-\beta t)\,dt$$

$^{+}$ Indicates SEER data that does not record a peak prior to the highest age category, which is mean 90 years.
$^{++}$ Indicates integration of SEER data up to age 90.

than the probability of the one random variable representing cancer extinction (see Appendix).

The Beta calculated peak incidence $I(t_p)$ and age at peak incidence t_p are compared to the SEER values. Several of the entries for the SEER data are noted to indicate that a peak was not recorded for those cancers in the age intervals reported, thus providing an uncertain SEER value for peak incidence and age at peak. For these cases it is best to refer to the figures to judge the adequacy of the estimate of SEER peak and age at peak. Since t_p is derived from the Beta function as $t_p = (k-1)/k\beta$, there is no dependence on α and only weak dependence on k. Accordingly, the age at peak incidence, which can be described as the turnover age, is almost entirely dependent on β, and as shown in the tables is consistent over all adult cancers and over a factor of 100 in incidence. The age at zero incidence represents the upper bound of the Beta function, and is equal to $\beta-1$. None of the SEER data extends to high enough age to test this model prediction, but cancers of the lung, larynx, brain, and corpus uteri show marked downturn of incidence within the age range reported. The Dutch and California data with older age groups show data tending to zero incidence at $\beta-1$ age.

The final column of each of the two tables present calculated cumulative probability of each cancer, based on the Beta fit, and assuming the individual lives to at least age $\beta-1$. The SEER comparison is the sum of the age specific incidence over all age groups. Since the SEER data does not extend to zero incidence, the SEER result should be somewhat lower than the Beta result, particularly if $\beta-1$ is higher than 90, which is the case. Individual cancer site probabilities rank approximately in order of the peak incidences, indicating that the incidence curve shapes are not too different from cancer to cancer, which can also be concluded from the constancy of k and β. For males, the maximum lifetime probability of an individual cancer ranges from 0.3% for Hodgkin's disease to 36.7% for prostate cancer. For females, the range is 0.2% for Hodgkin's disease to 20.7% for breast cancer. The calculated upper limit to the lifespan probability of any cancer for males at 70% and for females at 53%.

Comparison to Other Datasets

Figures 4(a-f) and Table 3 presents the de Rijke data for six major cancer sites, compared to the SEER data curve fits with the Beta function. Colorectal cancer incidence for Dutch males and females is not very different than the values for Americans, and closely agrees with the predicted turnover age and shape, even though the model was fit to SEER data that did not actually reach a peak (Figure 2b). Lung cancer shows different levels of incidence, but the location of the peak and curve shapes are similar to SEER. Of interest, the oldest male lung cancer group has incidence almost zero at age 100, which is close to the predicted $\beta-1$ value. Prostate cancer also shows about equal incidence and similar curve shape as the SEER fits, although the reported incidence appears to occur about 10 years later in the Dutch than in the SEER population. Female breast cancer appears to be a good match to the SEER fit curve shape at a somewhat lower incidence level, and suggests near-zero incidence at an age not too different from $\beta-1$. Bladder cancer and stomach cancer also show similar curve shapes to the SEER data fits, with the age at

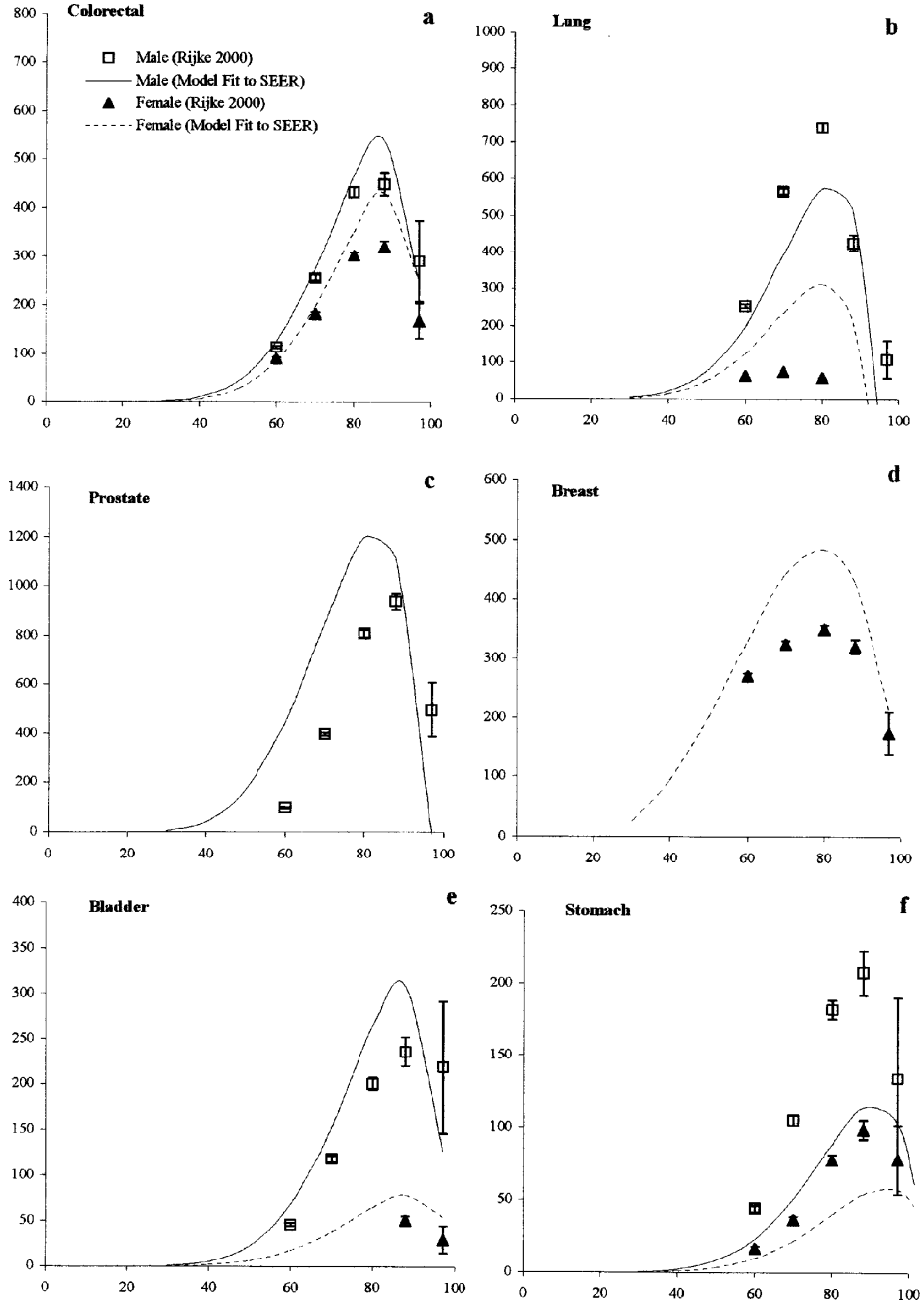

Figure 4a-f. Age specific incidence (per 100,000) vs. age data for Holland 1989-1995 (de Rijke 2000) compared to the SEER data fits with the Beta function for major cancer sites. Error bars indicate ± 2 SEM.

Table 3. Age-specific cancer incidence for major cancers in other countries compared to beta fits of SEER data: Holland* and Hong Kong+.

Site	Peak age-specific incidence (per 100k) Male Data; *[Beta]*	Age at peak incidence Male Data; *[Beta]*	Peak age-specific incidence (per 100k) Female Data; *[SEER]*	Age at peak incidence Female Data; *[Beta]*
Holland				
Colorectal	449; *[541]*	85-94; *[87]*	321; *[432]*	85-94; *[88]*
Lung	741; *[588]*	75-84; *[83]*	75; *[314]*	65-74; *[80]*
Prostate	939; *[1227]*	85-94; *[83]*		
Breast			349; *[486]*	75-84; *[79]*
Bladder	235; *[308]*	85-94; *[87]*	51; *[78]*	85-94; *[89]*
Stomach	207; *[117]*	85-94; *[91]*	98; *[59]*	85-94; *[94]*
Lymphomas	76; *[129]*	85-94; *[85]*	58; *[87]*	85-94; *[84]*
Hong Kong				
Bronchus, lung	827; *[588]*	80-84; *[83]*	427; *[314]*	80-84; *[80]*
Colon and rectum	437; *[541]*	85+; *[87]*	285; *[432]*	80-84; *[88]*
Bladder	224; *[308]*	85+; *[87]*	68; *[78]*	80-84; *[89]*
Stomach	221; *[117]*	85+; *[91]*	140; *[59]*	85+; *[94]*
Prostate	219; *[1227]*	80-84; *[83]*		
Breast			150; *[486]*	85+; *[79]*

*deRijke et al (2000), + Parkin et al (1997)

peak incidence correctly predicted. The model is particularly accurate in predicting stomach cancer peak, since the SEER data does not show a peak in its age range.

Figures 5(a-f) and Table 3 present Hong Kong age-specific incidence data for six major cancer sites, both male and female, with comparisons to the SEER model fits. Colorectal cancer incidence is about three-fourths of the SEER value, but the shape is similar, along with the age at peak incidence, to the SEER. Lung cancers are very close to SEER data in shape and age at peak incidence, with levels about one-third higher. Stomach cancer incidence is about twice that of the US, but appears to peak at about the same age. For bladder cancer, the incidence is lower than SEER for men in Hong Kong, but appears similar in age at peak incidence for both sexes. Conversely, prostate cancer is only one-sixth the SEER value, but with peak incidence appearing at about the same age. Breast cancer incidence appears quite different for Hong Kong women than their US counterparts, for reasons that are unknown.

Figures 6(a-d) present the Saltzstein *et al.* (1998) data compared to the Beta fit of the SEER data for six cancers. There is the expected good agreement for the age range up to about 90, which is the range reported by SEER. However, as we observed in the Dutch data, the turnover in incidence, and the continued decrease in incidence to age 100 predicted by the Beta model, is present. The slight rise in incidence for the oldest age group is ascribed by the investigators to be due to

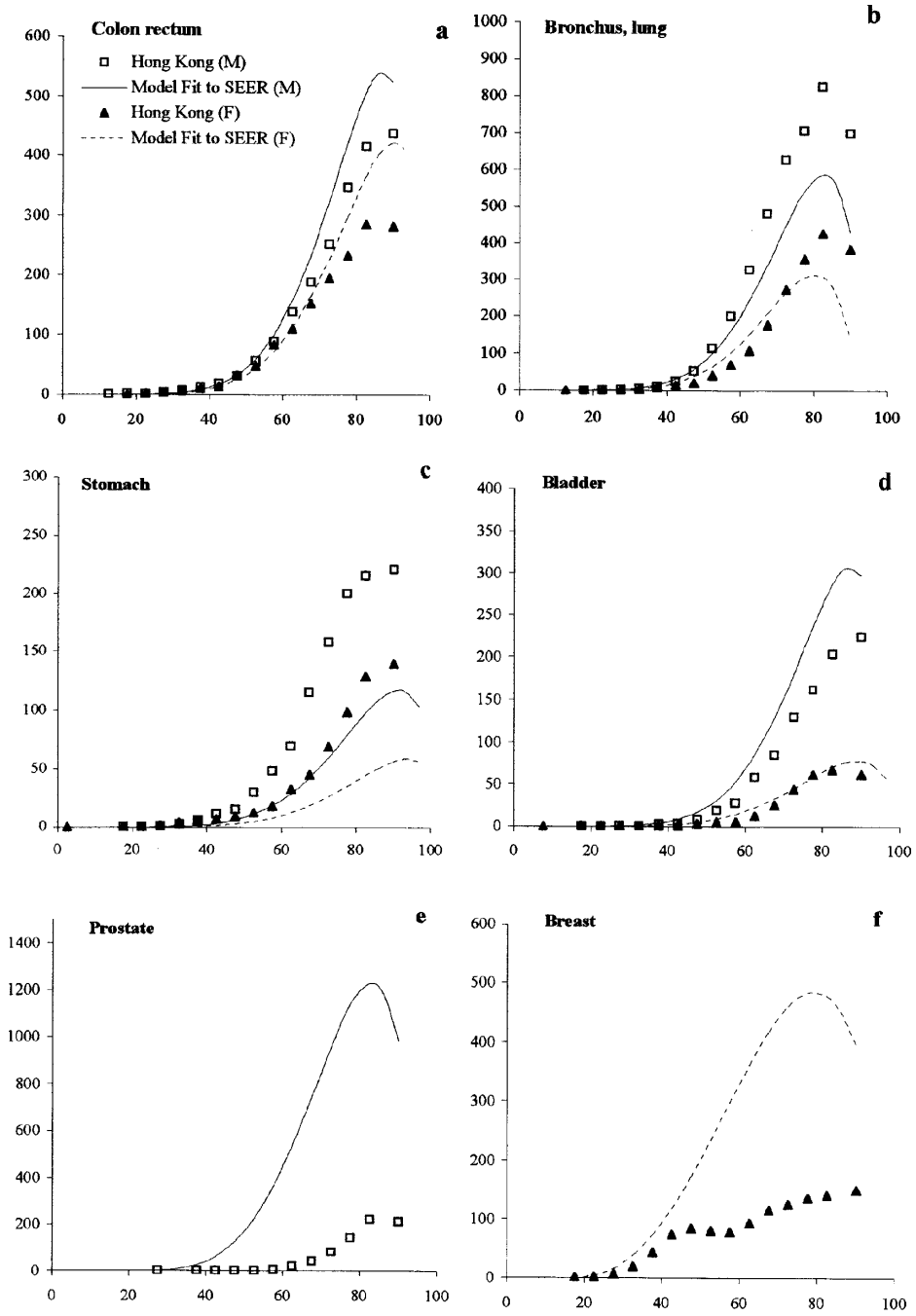

Figure 5a-f. Age specific incidence (per 100,000) vs. age data for Hong Kong 1988-1992 (Parkin *et al.* 1997) compared to the SEER data fits with the Beta function for major cancer sites.

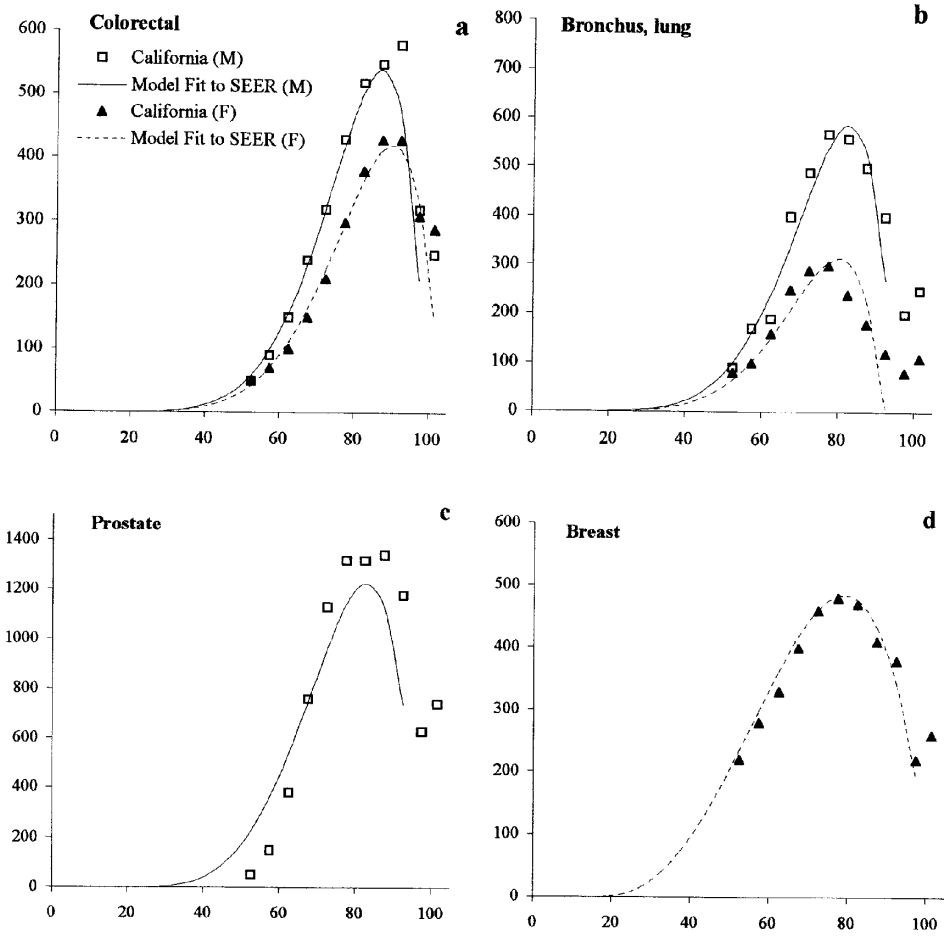

Figure 6a-d. Age specific incidence (per 100,000) vs. age data for California 1988-1993 (Saltzstein *et al.* 1998) compared to the SEER data fits with the Beta function for major cancer sites.

underreporting of the population of the ≥100 population over the relevant time period.

Comparisons of All Cancer Sites and All Populations

By normalizing each cancer vs. age data point to the peak value for that particular cancer and age group, we can plot the results on a single chart. Figure 7 shows all of the SEER incidences for adult male cancers (leaving out Hodgkin's disease, thyroid, testes) plotted together, along with the mean value of all of the SEER incidences at each age. The Beta model fit to the SEER data is included, extending to age 101. Also plotted are the Dutch, Hong Kong and California incidence data. The Dutch and California data are particularly valuable because they extend to age 97 and 102, respectively, where the SEER data ends at 90. By inspection of Figure 7, all of the male adult cancers incidences fall into a well-defined band, despite the

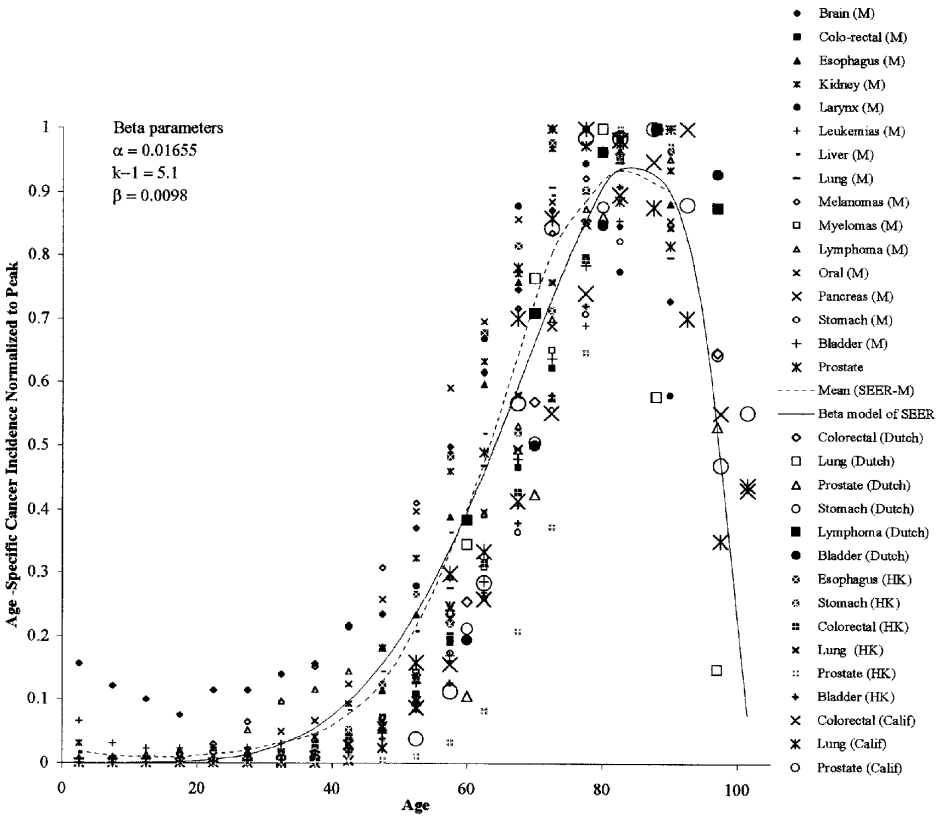

Figure 7. Cancer incidence vs. age for all SEER male sites except for childhood cancers (Hodgkin's, thyroid, testes). Each incidence is normalized to the peak value for that specific cancer. Included for comparison are the data for Dutch, Hong Kong, and California male sites, and a Beta fit of the SEER data.

factor of 100 variation in peak incidence for the range of cancers considered. The band scatter standard deviation averages approximately ±8% of the peak incidence about the mean at each age group of each cancer.

DISCUSSION

There are five alternative ways of describing the modeling of the cancer incidence data (all of them assuming the data at elevated age are valid, a point discussed further later):

1. The simple Beta function fits the age distribution including the turnover at elevated age well, while previous biologically based models have been unable to do so, which in turn leads to a search for a biological basis for the Beta model.

2. The curve shape for adult cancers, including the turnover, appears consistent from male to female, from culture to culture, and even from cancer to cancer, varying only in level of incidence.

3. There is apparent remarkable uniformity of the age at peak incidence across all adult cancers, despite a factor of 100 difference in peak incidence in these cancers.

4. Extrapolation of the Beta function fits beyond the age for which there are data allows us to calculate the age at which incidence is expected to be zero. Then we may integrate the incidence to calculate a lifespan cumulative probability for each cancer, and all cancers combined.

5. The cumulative probability of a person contracting any cancer is less than one, and of each individual cancer it is much less than one. The conventional wisdom that everyone will eventually contract any or a specific cancer if he or she does not die of some other cause, may be incorrect.

Curve Shape: Comparison to Other Models

The earlier models directed to explaining age distribution of cancer are of two general types: multistage, and clonal expansion. The earliest derivations of the multistage view approximated the model as the product of independent probabilities of stage transitions, $\mu_1 t \ldots \mu_{k-1} t$. Specifying the order of the transitions resulted in the age distribution of cancer incidence as $I(t)=t^{k-1}(\mu_1\mu_2 \ldots \mu_k)/(k-1)!$, where μ_i are the transition rates for each stage, a result first proposed by Armitage and Doll (1954). This form we refer to as the A-D power law model. Although highly successful in fitting the rising side of cancer incidence data (up to age 74), it is obvious by inspection that this model cannot fit the turnover, and thus cannot produce the desired shape. Moreover it became clear that the formula was only an approximation, valid only for low cancer rates. A mathematically exact form is discussed later.

The clonal expansion model is based on the hypothesis that no more than two stages were supported by biological evidence, and that a cell need undergo only two transitions to become malignant, with the first transition conferring a survival advantage causing exponential growth of the cell by division. The second transition is required in order to release the cell from control completely and become malignant. Accordingly the incidence may be approximated as $I(t)=\mu_1\mu_2 e^{bt}$, where μ_1, μ_2 are the rates of the two transitions, and b a growth factor, as Armitage and Doll noted (1957). Although useful for fitting incidence at small values of t, this model, which we can refer to as the simplified A-D clonal expansion form, cannot produce the age turnover. A more complete derivation results in $I(t) = N\mu_1\{1-\exp[-\mu_2(e^{at}-1)/b]\}$, where N is the mean number of cells per person exposed to the first transition. This incidence function increases monotonically and approaches $N\mu_1$ as $t \rightarrow \infty$, thus avoiding the fate of limitless growth of incidence in the simpler A-D clonal expansion expression, but it likewise cannot produce a turnover.

By adding additional features to the clonal expansion model: that the number of cells at risk might be a variable, and that transformed cells have a death rate as well as a proliferation rate, the incidence may be approximated as: $I(t) \approx \mu_1\mu_2 \int N(s)\exp[(\alpha_2 - \beta_2)(t-s)]ds$. The integration is taken from 0 to t, α_2 and β_2 are growth and death rate of transformed cells respectively, and $N(s)$ is a variable cell number function. Holding $N(s)$ constant, we integrate to: $I(t) = \mu_1\mu_2 N\{\exp[t(\alpha_2 - \beta_2)]-1\}/(\alpha_2 - \beta_2)$, which

produces a convex monotonically increasing exponential curve if $\alpha_2 > \beta_2$, or a concave asymptotically limited curve if $\alpha_2 < \beta_2$ as $t \to \infty$. This approach, well known as the simplified MVK model, was developed by Moolgavkar and colleagues (1981) and has been very successful in modeling many cancers. It is clear the model can produce an age turnover by applying a suitable function $N(s)$, and specifying that $\alpha_2 < \beta_2$, which Moolgavkar *et al.* (1981) proposed. Although quite successful, this simplified form is known to have limitations and therefore the exact form of the two-stage clonal expansion model is currently recommended, although it requires numerical procedures, with no closed form of solution readily accessible (Moolgavkar *et al.* 1999).

As shown in the Appendix, the Beta model results from adding to the A-D power law model, the probability of a cancer extinction step, which is modeled as a uniformly distributed random variable over the interval $(0, \beta t)$. Applying the same cancer extinction step to the simplified MVK clonal expansion model with constant parameters might fit the cancer incidence data with turnover as well as the Beta model, as can be inferred from Figure 1. Although the simplified MVK form already includes an explicit factor $(\alpha_2 - \beta_2)$ modeling the difference between birth and death rates of initiated cells, current thought is that this deterministic approach is incorrect since the initiated cells appear to have a stochastic character: the probability of initiated cells being present in the tissue and consequently the probability of cancer per unit time, is greater than zero even if $\alpha_2 < \beta_2$ for long times (Moolgavkar *et al.* 1999 p.197). This stochasticity assumption is an important part of the exact form of the MVK model. The Beta model prediction that the probability of cancer per unit time goes to zero with certainty at $t \geq \beta - 1$ is clearly different.

Age at Peak Incidence: Comparisons to Other Models

The tabulations of age at peak incidence of Tables 1 and 2, derived from the Beta fits, also evident in Figure 7, are quite uniform for the adult cancers: male 85.0 mean $\pm$ 3.7 s.d., and female 84.5 $\pm$ 7.1 (the s.d. indicating the standard deviation of the age at peak incidence over all cancers). The Beta distribution formula for the age at peak incidence, $t_p = (k-1)/k\beta$, has no dependence on the cancer creation coefficient α , is only weakly dependent on the number of stages k, and is almost entirely dependent on the value of the cancer extinction factor β. Earlier models have not produced this constancy of t_p , and when tested tend to predict a much different result as discussed below.

Armitage and Doll (1954) fitted a power law to age-specific mortality data to age 74, which they assumed was a good representation of age-specific incidence since cancer victims quickly died, and found $I(t) = at^{k-1}$. This fit should be interpreted as a hazard function $h(t) = f(t)/[1-F(t)]$: the probability of dying at age t (a probability distribution function, pdf, which is $f(t)$), is conditioned on survival to age t (one minus the associated cumulative distribution function, cdf, which is $F(t)$); since the victims' death removes them from the denominator when computing the ratio of cancer deaths to population at risk. We denote age-specific mortality $m(t)$ to indicate this hazard function. For individual cancer mortality data, the cumulative probability of death from any specific cancer $F(t)$ by end of normal life is of order a few percent, and the A-D approximation that the incidence $I(t) = m(t) \approx f(t)$ is quite good.

It is only when considering the turnover in $I(t)$, which must occur to the pdf by the unitarity criterion (must integrate to one), do we need consider the exact pdf expression derived from the data fits.

Accordingly we write $I(t)=m(t)=at^{k-1}= f(t)/[1-F(t)]$ exactly, and note that for small values of cancer cumulative probability $F(t)$, $I(t)=at^{k-1} \approx f(t)$, which is the usual approximation taken, resulting in a pdf that appears to grow without limit. However to consider age at peak incidence we must consider high values of $F(t)$, and thus use only the exact hazard function implied by the A-D power law fit: $at^{k-1}= f(t)/[1-F(t)]$, to derive the exact pdf implied by that power law fit.

Rewriting the hazard function in its differential form and integrating, we obtain the exact probability of cancer based on the fit as $F(t)=1-exp[-\int h(t)dt]=1-exp[-\int at^{k-1}\, dt] =1-exp[-(at^k)/k\,]$ (integration from 0 to t), which clearly gives $F(t)= 1$ as $t\rightarrow \infty$, and thus is a cdf (this is so for any $h(t)>0$). Differentiating with respect to time, we write the pdf associated with the cdf as: $f(t)= (at^{k-1})\, exp[-(at^k)/k\,]$. Noting that $f(0)= 0$ *and* $f(\infty)= 0$, there is a peak in $f(t)$, which we derive as $t_p = [k(k-1)/a]^{1/k}$. Since the incidence level at any age is determined by a and k, and $k \sim 7$ for most cancers, a variation of a factor of 100 in incidence from rare cancers to common cancers implies a shift in the value in t_p of a factor of two, which is a very different result than for the Beta model and the SEER data.

The exact pdf derived from the A-D power law hazard function fit to age-specific mortality data, is the unconditional age-specific incidence, which would be measured if all cancer victims remained in the population. This is based entirely on the assumption that the mortality data is accurately fitted by at^{k-1}. This is clearly not the case above age 75, but its failure to fit at high age does not appear to be due to mathematical approximations, but due to additional biology not modeled by the power law.

Armitage and Doll (1954) inferred a detailed multistage model for cancer from their biological interpretation of the power law fit (and other evidence), for which the power law model is an approximation. Their model can be made mathematically exact by solving the system of differential equations describing the probability of finding a cell in each of the stages in its transitions to cancer. Moolgavkar (1978, 1999) found a method of expressing this exact pdf as a MacLaurin series expansion as: $f(t)=[t^{k-1}(\mu_0\mu_1\ldots\mu_{k-1})/(k-1)!][1-\mu t+f(\mu,t)]$, where μ_i are the transition rates for each stage, and μ is the mean of the transition rates. It is important to note that the above expression was derived as a pdf, and the approximation taken that $I(t)\approx f(t)$, valid for small values of $I(t)$, as was assumed by Armitage and Doll for the power law model. The reader is carefully alerted to the fact that there is a difference between epidemiological data inferring hazard function, and theoretical derivations inferring pdfs.

We can immediately note that the first term is the A-D power law as a first order approximation to the multistage model, *or* the exact hazard function of the power law fit. Adding the second term results in an expression that is very similar to the Beta function, but with constants that are not arbitrary. Since this derivation is based on the exact multistage pdf, and there is similarity of form to the Beta function, we might investigate its properties further.

Assuming that the two terms are adequate to test the model for its prediction of the shift of age at peak incidence with incidence level, we can then derive: $t_p=(k-1)/$

(μk). Since incidence is proportional to the product of the transition rates μ_i , and if those rates are all approximately equal such that μ varies with incidence to the k^{-1} power, and $k \sim 7$, it is easy to see that that t_p shifts by a factor of two for 100-fold change in peak incidence. However, this two-term expansion form of the exact multistage model leaves open the mathematical possibility that large changes in incidence may be produced by making one or more μ_i very much smaller than the others. The average of the μ_i then becomes constant, while permitting unlimited change in incidence by changing the small μ_i, thus making t_p constant. Accordingly we cannot rule out this model from also producing the SEER data fits observed with the Beta function, but observe that all of the μ_i for all adult cancer sites (probably in all countries) must conspire to produce exactly the same average value, within a few percent, while producing peak cancer incidences for each of those sites that vary over a factor of 100.

The Beta distribution model is, as are the A-D and Moolgavkar models, derived as a pdf, $f(t)=(\alpha t)^{k-1}(1-\beta t);\ 0 \le t \le \beta^{-1}$, with the approximation taken that the pdf is a good model of the SEER incidence data: $I(t) \approx f(t)$. This is clearly accurate for small values of cumulative cancer probability $F(t)$, but leads to the question as to whether the SEER data should be considered a hazard function (all people with that cancer removed from the denominator) or a pdf (all people with that cancer remain in the denominator) when modeling high values of $F(t)$ at the turnover age of common cancers (lung, colorectal, prostate, and breast cancers have cumulative probabilities greater than 10%: see Tables 1 and 2). Since the overall mortality rate from cancers in the SEER data is about one-half of the overall incidence, the SEER data suggests an interpretation about midway between a hazard function and a pdf, *i.e.,* about halfway between $I(t) \approx m(t)$ and $I(t) \approx f(t)$.

One possibility is to carefully account for the survival fractions for each cancer for each age group and construct a fit to $f(t)=[I(t)][1-M(t)]$, where $M(t)$ is the cumulative number of people to die of the cancer. For example, mortality is high for lung cancer, thus $M(t) \sim F(t)$, and prostate cancer mortality is low, thus $M(t) \sim 0$. It should be noted that for prostate cancer, the high cumulative incidence with low mortality would tend to increase the fraction with the cancer in the population in the SEER data, thus reducing the pool without prostate cancer, and causing a turnover in reported incidence as the cumulative incidence approaches unity. However maximum cumulative incidence is only 37%, which too far from unity to cause the marked turnover observed, particularly in the Dutch and California data.

The Beta distribution proves to be very robust when modeling data with uncertain removal by death, giving the same results in curve shape, quality of curve fit, and constant t_p, with either $M(t)=F(t)$ or $M(t)=0$ interpretation of the data. This observation results from writing the exact hazard function $b_h(t)=b(t)/[1-B(t)]$, where $b(t)$ is the Beta distribution, $B(t)$ its integral, and $b_h(t)$ the hazard function associated with the Beta distribution. Then $b_h(t)= [(\alpha t)^{k-1}(1-\beta t)]/[1-(at)^k(1-bt)]\ ;\ 0 \le t \le \beta^{-1}$, where $a=[\alpha/k^{1/(k-1)}]^{(k-1)/k}$ and $b=k\beta/(k+1)$. Since $b<\beta$, then $b_h(t) \to 0$ *as* $b(t) \to 0$, thus predicting the identical age at zero incidence, which is a critical feature of the model. If the SEER data is fitted as $I(t)= b_h(t)$, then the parameters α, β, and $(k-1)$ will change slightly, but produce the same curve shapes with same fit quality, and the same age at peak incidence. Accordingly, we can conclude that the Beta distribution, $b(t)$ models the

incidence data in a robust way, and is not sensitive to mortality rates for its major features.

Extrapolation of the Beta Distribution Fit

Since the Beta distribution is a successful fit we venture to extrapolate the distributions beyond the turn over where data are non-existent for the SEER data or limited for the Dutch and California data. Figure 7 suggests that all adult cancers might share a uniform characteristic of power law or exponential growth at about the same rate to about age 70, where the incidences level off and eventually reduce toward zero at age ca. 100. Accordingly, the Beta fit equation: $I(t)=(\alpha t)^{k-1}(1-\beta t);\ 0\leq t\leq \beta^{-1}$, with $\alpha=0.01655,\ (k-1)=5.1,\ \beta=0.0098$ provides a useful general formula for the age distribution of any adult cancer as a fraction of its peak value. When considering absolute values of incidence, the cancer creation coefficient α scales the curve to the appropriate level.

The extrapolation of the beta distribution yields the interesting parameter, which is the age at predicted zero incidence, which is simply $t_0 = \beta^{-1}$. This discussion is clearly more speculative, but if we make the obvious interpretation, after about age 100 cancer incidence falls to zero. There is general agreement in the literature that cancer is a less threatening disease for persons living to age near 100 (Smith 1996, Stanta 1997, Saltzstein 1998), but the Beta prediction that cancer incidence (both the pdf and hazard function) will fall to zero with probability one is new. These observations suggest a new view that is different from the conventional wisdom, which was largely based on the historically important models described above: that cancer probability continues to increase with age until it reaches certainty.

Cumulative Cancer Probability

Tables 1 and 2 show the cumulative probabilities, calculated from the beta distribution fit $I(t)=b(t)$, of each cancer and all cancers over a lifespan (defined as surviving to $age>\beta^{-1}$). For males, these range from 0.3% to 31% for individual cancers, and 70% for at least one cancer of any type. For females the range is 0.2% to 20% for individual cancers, and 53% for any cancer. The simple and obvious conclusion is contrary to the common understanding that: "if a person lives long enough he or she will get cancer," which is a result of the historical success of the simple power law and clonal expansion models, both of which imply a rising probability that always reaches unity at large enough values of t.[1] The data, as interpreted with the Beta model, suggest that "if a person lives long enough, he or she may avoid cancer entirely," with about a one in three chance for men and an even chance for women.

Modeling Susceptibility and Sensitivity

Both the multistage and clonal expansion hazard function models, whether approximations or exact, have the characteristic that the pdf of any cancer integrates to one over sufficiently long time $\{\int f(t)dt=F(t)=1-exp[-\int h(t)dt]=1$, *integration limits* $0\rightarrow\infty$ *; valid for any positive function* $h(t)\neq 0$ *as* $t\rightarrow\infty\}$. As the data (including the

[1] This applies also to the mathematically more exact versions of the multistage model.

extrapolation with the Beta function) indicate, however, the cancer incidence yields cumulative probabilities much lower than one, and range over a factor of 100, while maintaining similar curve shape. One simple and obvious assumption is that only a fraction C of the population is susceptible, which leads to modification of the Beta function model as $b(t) \approx I(t) = C(\gamma t)^{k-1}(1-\beta t)$; where $C = \int(\alpha t)^{k-1}(1-\beta t)dt$, and $\int(\gamma t)^{k-1}(1-\beta t)dt = 1$; $0 \leq t \leq \beta^{-1}$. Numerically, the susceptibles fraction is identical to the cumulative probability over lifespan tabulated in Tables 1 and 2. Inherent in this interpretation is that the fraction of susceptible people is different from one cancer site to another by about a factor of 100.

Cook *et al.* (1969) added a limited pool of susceptibles expression to the A-D power law, which produced a turnover, but they found that the location of the age of peak incidence varied markedly with the incidence level. Since the data did not support variation in the age at peak incidence, they deemed this hypothesis unsupported. Herrero-Jimenez *et al.* (1998, 2000) employed a biologically detailed modified clonal expansion model to examine colon cancer mortality turnover, which might avoid the Cook problem with their hypothesis of an exposure factor in addition to a susceptibles factor. This allows the shape to be held constant with one factor, with the level controlled by the other factor, but relies on the same assumption: turnover occurs because we "run out of candidates" beyond about age 80.

Finkel (1995) raises the importance of distribution in susceptibility in risk assessment and formulates an interesting analytical method of modeling susceptibility by combining a lognormal distribution assumption with a modified Armitage-Doll cancer model. He shows that including the susceptibility distribution assumption causes the modeled age-specific incidence to plateau at an elevated age, thus improving the fit to colon cancer mortality data compared to an unmodified A-D model. The biological basis for a distribution in susceptibility is certainly plausible, given the heterogeneity in genetic, environmental, and life style influences on cancers. Finkel clearly supports the idea that the data indicating flattening at old age is not artifactual, but like the Cook and Herrero-Jimenez models, the flattening occurs in his model because the susceptibles pool is being depleted. However, Finkel's model cannot predict an actual decrease in cancer incidence until the cumulative incidence approaches unity, which no individual cancer approaches.

The weight of the evidence seems to argue against a distribution of susceptibles view, since the distribution would have to be quite similar for each of the 35 adult cancers to peak at close to the same age over an incidence range of a factor of 100. This suggests a biological mechanism, which is uniform in its genetic or environmental influence, opposite to the Finkel view that requires heterogeneity. Additionally, in a study performed in parallel with this work, Pompei *et al.* (2001) analyzed animal data for a single species of inbred mice living their lives in a controlled uniform environment, which show similar curve shapes and turnover in cancer incidence. Interestingly the turnover for mice also occurred at about 80% of the lifespan. These new observations clearly tend to weaken the susceptibles view based on heterogeneity, and strengthen the view that a biological process, yet to be modeled, must be considered.

[2] Thus it is not in contradiction to the observation that 100% of the highly exposed β-naphthylamine workers developed cancer.

A further consideration for the susceptibles hypothesis is data on persons heavily exposed occupationally. In one well-documented case, exposure to β-naphthylamine, 15 out of 15 persons exposed developed bladder cancer (Case *et al.* 1966).) These data suggest that everyone is susceptible to cancer if the dose is suitably high, although variations in basic sensitivity by a factor of 3 are possible, as discussed by Finkel. The Beta model does not preclude rates of cancer approaching unity at high doses, but does require that they occur at younger ages, where the age-related slowing of the cancer process suggested by the $(1-\beta t)$ term is less important.[2] The Case *et al.* paper is instructive on this point, showing convincing data that chemical workers with high bladder cancer rates contracted the cancer at ages 20 years younger than the general population contracted the same cancer. This observation suggests that susceptibility and exposure might be a valid method of modeling the rising side (and perhaps flattening, as Finkel suggested) of the age distribution. However, at age greater than about 80% of lifespan, the SEER, Dutch, Hong Kong, and California data suggest a uniform cancer extinction biology may dominate.

Decreased stage sensitivity at old age might produce both a flattening of the cancer incidence and a turnover if sensitivity approaches zero. Consider the A-D multistage model for a relatively uncommon cancer such as liver cancer (cumulative lifespan incidence ~ 1%), where the A-D approximation is accurate: $I(t) = at^{k-1}$. If the stage probabilities are not equal and constant, the A-D model becomes $I(t)= (p_1t)(p_2t)...(p_{k-1}t)p_k/(k-1)!$, where (p_it) are the transition probabilities for each stage. Since this model includes the requirement that the stages $(1,2,\ldots,k)$ must occur in order, then it is the later stages that are of interest, since the earlier stages have already occurred if they are going to, and a change in early stage probability at larger t will not alter the overall probability (Armitage and Doll 1954). By inspection, it is clear that if p_{k-1} or p_k approaches value *0* at some time t approaching *age ~100 years*, then $I(t)\rightarrow 0$, and thus produces a turnover. Thus if a decrease in sensitivity is interpreted as a reduction in probability of a late stage transition, a flattening will be produced, followed by turnover as the probability of the late stage approaches zero. A similar argument can be made for the clonal expansion models. Accordingly, the Beta model cancer extinction factor might be equally interpreted as a linear sensitivity decrease with age of a late stage: $p_k = \mu_k(1-ct)$, with no loss of generality or goodness of fit.

Other forms for this sensitivity reduction factor, such as e^{-ct} might appear also to work adequately (since the first two terms of its expansion are also $(1-ct)$). This produces a Gamma function form $(\alpha t)^{k-1}e^{-ct}$ when combined with the A-D power law, and avoids the slight mathematical discomfort of negative incidence when $t>\beta^{-1}$ with the Beta form. However the fit is not nearly as satisfactory as the Beta function form, and the values of the constants become seemingly unrealistic when as good a fit as possible is forced. For example, $(k-1)$ is about 5 for the Beta, but is about 15 for the Gamma to fit the SEER data, suggesting an unrealistically large number of stages. Also the extinction coefficient, β, has to be much larger in the Gamma, and seriously distorts the fit at low values of age. We conclude that the Beta form, although having an abrupt limit at $t=\beta^{-1}$, nonetheless is what the data suggests. The exponential form of extinction factor might, however, work well if applied to a different cancer creation model, particularly an exact form, but this has not yet been explored.

Biological Hypotheses

The fact that the Beta distribution fits the data well, and that the multistage and clonal expansion models appear not to do so, even when made mathematically exact, suggests that we call the Beta distribution a "model" and enquire about its possible biological plausibility. The Beta model suggests a very different cause of the turnover: the active involvement of a cancer extinction step such that the probability of a cancerous cell survival (or proliferative ability) approaches zero at ages corresponding to approximately a human lifespan. The model can be derived from the first order multistage model (power law fit hazard function) by adding a "cancer extinction factor" $(1-\beta t)$ by which the transformed cells are eventually destroyed or deactivated at a rate greater than their creation (derived in the Appendix). However, it might be incorrect to interpret the constant β in the fits as this factor, since an exact multistage or clonal expansion model with cancer extinction might have a somewhat different formulation.

The Beta model applied to the biology can be viewed as a simple combination of two factors: (1) cancer creation, which is most simply modeled with a power law multistage assumption, although it would fit equally well with an exponential clonal expansion assumption or most any other rapidly increasing function; and (2) cancer extinction, which is modeled as a cumulative probability that linearly increases to certainty at age ~100. The first factor may be interpreted in the same way as all of the relevant historical models: caused by mutations and promotion steps from genetic, environmental, etc. exposures. The extinction factor is new, and its biology must be carefully considered.

Commonly accepted, but not entirely understood, inexorable changes due to ageing might lead to clues. As a first possibility, apoptosis is a candidate for the mechanism of "cancer extinction". Although we added this term by assuming a process that is uniform with age, an age dependence might also be included. *In vitro* human cell studies by Schindowski *et al.* (2000), Lechner *et al.* (1996), and Potestio *et al.* (1998) found that apoptosis increases with age due to reduced defense from oxidative attack. Higami *et al.* (2000) suggest apoptosis increases *in vivo* with level of accumulated injury related to ageing. Ogawa *et al.* (2000) found apoptosis rates low in the young, and increased in the old from bone marrow samples from newborns to age 100. Lee *et al.* (2000) found that rat colon epithelial cells were more sensitive to apoptosis stimulation with advancing age.

A second possibility is cell senescence, or loss of proliferative ability, which may be interpreted as a loss of sensitivity. This point has been discussed by Faragher (1998, 2000) who suggest that cell senescence, like apoptosis, occurs as an anti-cancer mechanism, and that a large body of evidence suggests cell senescence contributes to a variety of pathological changes seen in the aged. Hayflick (2000), Jennings *et al.* (2000), Oloffson *et al.* (1999), and Rubelj *et al.* (1999) all suggest cell senescence or the related observation of telomere shortening increases with age, and thus may profoundly influence the cancer process. Rubelj further raises the interesting possibility that telomeres may shorten abruptly by a stochastic process, thus producing senescence in some cells even at young age.

If the probability of abrupt shortening were uniform with time, this mechanism could be modeled exactly as causing cell senescence with probability of βt, and thus

the cancer extinction factor becomes the $(1-\beta t)$ proposed in the Beta model. Such a process is suggested by observations beginning decades ago showing that replicative ability of cells markedly decreases as they age, which ultimately defines senescence (Hart *et al.* 1976, 1979). Further, the loss of replicative ability appears to reduce approximately linearly with age, thus suggesting a factor such as $(1-\beta t)$. This approach is proposed and discussed in Pompei *et al.* (2001).

Data Reliability

The conclusions of this work rely critically on the assumption that the modern cancer registry data is truly representative of cancer incidence, particularly above age 80. The concern first expressed by Armitage and Doll (1954), that less extensive workups are performed in diagnosing cancer for older persons than for younger, cannot be completely dismissed. However, there are accumulating evidence via autopsy of the oldest old that cancer prevalence indeed reduces with increasing age. The Stanta autopsy study mentioned earlier, which included 507 people who died between the ages of 75 and 106, was designed to investigate this very question. The authors state "Ö our autoptic population may be considered representative of the general population" and "We discovered a cancer in 36% of the people between 75 and 90 years of age, but only in 22% of those over 95, and in 16% of the centenarians." The details of the histologic examinations were not reported. However the authors do find substantial and increasing underreporting of cancer with age when the autopsy results are compared to the original clinical diagnoses. Imaida *et al.* (1997) studied autopsies of 871 patients aged 48 to 113 at death and also found that prevalence of malignancies reduce at the older ages, but also found increasing prevalence of latent cancers with age, latent cancers defined as those not diagnosed clinically.

The de Rijke study reported histologically or cytologically confirmed cancer diagnoses in 98% of males and 97% of females in the 55 to 64 age group, and 87% and 84% for those in the ≥95 age group, suggesting the possibility of a reduction in thoroughness for the older ages. Referring to Figure 1, we see that without some important effect(s) at age > ~75 (these effects might be depletion of susceptibles, increased apoptosis, increased senescence, slowing of proliferation, and/or underreporting of cancers), cancer incidence is expected to continue to increase strongly, by the historical paradigm, until the cumulative incidence reaches unity. At about age 75, 10 to 15% deficit in cancer diagnosis might be sufficient to account for the deficit between expected and observed incidence. At age ≥ 95, however, deficits of a factor of two or more from the aforementioned effects are required to reduce incidence to the observed values. Referring to Figures 4 and 5 for the Dutch and California data for individual cancers, we observe that the deficit in incidence, compared to even a conservative straight line trajectory extended from the 60, 70, 80-year-old incidence rates, is a factor of two or much more by age 95.

SEER have not specifically addressed the issue of the reliability of the cancer incidence reporting for the oldest age groups, but believe the data is at least as reliable as that reported by other countries (Ries 2001 personal communication). SEER themselves seem to accept the data are reliable enough to describe a turnover, and further have observed, "Whatever is occurring, fortunately cancer is not inevi-

table for all older persons" (Yancik and Ries 1995). We have also not considered other possible influences such as altered diet, lifestyle or environment for the oldest, which may tend to reduce cancers by mechanisms other than age, which suggests further study. Although we are not yet able to rule out concern about cancer reporting (or lifestyle effects) in the oldest, the weight of the evidence, including the previously discussed mice data, is tending to support the validity of the reported data.

Future Work

The questions raised by this work are centered ultimately on the compelling possibility of outliving our cancers. From epidemiology, is there evidence that as population longevity increases, that cancer incidence naturally reduces, independent of prevention measures? Mathematically, this question seems the same as choosing whether incidence reduces because a pool of susceptibles is depleting, or that there is a mechanism that arrests cancer development at elevated age, that might be independent of an individual's lifespan. The latter might be clarified by examination of the various exact formulations of the cancer models, which all in some way must include depleting a susceptibles pool with age. If such models do not fit the data, does addition of the cancer extinction factor produce a fit?

The specific areas that might be considered are

1. Improve our confidence in the reliability of the data particularly above age 80, and the variation of that reliability with age, and include possible birth cohort effects.

2. Compare the Beta model fits to more data sets, again particularly above age 80 and in registries with historically reliable data, including detailed analyses of the unpublished age-specific SEER data to age 100.

3. Explore a more exact form for the Beta fit by including the mortality data in the hazard function interpretation of the SEER data.

4. Investigate results of animal bioassays for evidence of the incidence turnover, particularly for bioassays without terminal sacrifice.

5. Add a cancer extinction factor to the exact multistage and clonal expansion models, exploring the fits and biological implications.

6. Search for evidence of a cancer extinction effect in dose-response and other aspects of carcinogenesis, particularly as dose might influence a biological effect modeled parametrically as β.

ACKNOWLEDGMENT

The authors thank Dr. Suresh Moolgavkar of the Fred Hutchinson Cancer Research Center, Sir Richard Doll of the University of Oxford, Dr. James Wilson and

Dr. Dan Kammen both of UC Berkeley, and Dr. Chiu Weihsueh of the USEPA for all of their helpful comments.

APPENDIX

The selection of the Beta distribution for the data fits arises from the observation that the power law equation $I(t) = at^{k-1}$ well fits many cancer site incidence data up to about age 74 (ignoring childhood cancers) At older ages, the incidence data markedly flatten, and show reduction at sufficiently elevated age. Accepting the validity of the power law fits at younger ages (but not necessarily the validity of the power law model itself), we add the hypothesis that a "cancer extinction" term is influencing the carcinogenesis process, eventually becoming dominant at sufficiently elevated age.

Adding this cancer extinction term to the power law is accomplished directly by forming the probability statement: *probability of cancer = the probability of reaching k stages and the cancerous cell does not die (or lose its proliferative ability).* We write this probability and expand as: $P_c = P(bt^k \cap \text{not death}) = P(bt^k | \text{not death})P(\text{not death}) = bt^k * P(\text{not death})$. The simplest assumption for a probability density for cancerous cell death is a uniform distribution over the time interval 0 to ct, leading to $P_c=bt^k(1-ct)$, where c is a constant. Taking the time derivative to convert the probability to a probability density function for a single cell: $f(t)=\alpha t^{k-1}(1-\beta t)$, where α and β are constants. We immediately recognize the Beta distribution $f(x)=\lambda t^{r-1}(1-x)$ over the interval $0 \leq x \leq 1$, where $x=\beta t$. A textbook interpretation of $f(x)$ is the density for the $(r-1)^{th}$ largest of r uniform (0,1) random variables (Larson 1982), which can be restated as the probability density function for achieving $(r-1)$ stages (cancer creation) without achieving the r^{th} stage (cancer extinction).

Expanding from consideration of a single cell to N cells in an organ, and denoting $f(t)=F'(t)$, the probability of cancer is $G(t)=1-[1-F(t)]^N$. For large N, this simplifies to $G(t)=1-e^{-NF(t)}$, which is accurate to 10^{-10} for $N=10^8$ cells. As discussed by both Moolgavkar (1978) and Armitage (1985), the age-specific incidence function for the organ tissue is not the density function $G'(t)$ itself, but the associated hazard function, given by $h_c(t)=G'(t)/[1-G(t)]$, which represents the incremental risk of cancer per unit time given that the tissue has been cancer-free to time t. Completing the derivation, $h_c(t)=e^{-NF(t)}\ Nf(t)/\ e^{-NF(t)} = Nf(t)$. We note that the age-specific cancer incidence for a site tissue is related to the probability density function for one cell by the constant N, thus leaving the Beta model as $f(t)=\alpha t^{k-1}(1-\beta t)$, modified by only by a constant (absorbed into α) to apply to a multicellular organ site. The final expression chosen immerses the α constant into the $k-1$ power in order to preserve the historical view of $k-1$ stages, each with its own transition rate (assumed to be equal in this case), thus denoting the final form as $b(t)=(\alpha t)^{k-1}(1-\beta t)$.

To apply the Beta model to fit epidemiological age-specific incidence data for a specific cancer,' $I(t)$, we consider whether the data is properly interpreted as a hazard function or a pdf. Since the hazard function is given by $I(t)=f_e(t)/[1-F_e(t)]$, the subscript e denoting a pdf and cdf derived from epidemiological data, which is the number of new cancer cases divided by the population at risk; the question reduces to whether the data set modeled has in the denominator only the population at risk for that cancer or the entire population for that group. For the SEER data, the

denominator includes all members of an age group still alive at time t, which includes all who have been diagnosed with a cancer still alive. If the mortality due to cancer were zero, then the usual approximation $I(t) \approx f_e(t) = b(t)$ (valid for small $F_e(t)$) would be exactly $I(t) = f_e(t) = b(t)$. Since mortality rate is about one-half of incidence rate overall for the SEER data, the exact statement cannot be made, and the approximation must be taken. It should be noted that since only one-half of the $F_e(t)$ are removed from the denominator, this approximation is considerably more accurate than if incidence is inferred by age-specific mortality in which the approximation is taken that $I(t) \approx f_e(t)$. Further discussion of this point is in the main text.

As employed for the fits, the Beta model does not integrate to 1, as a correct density function must, but integrates to the cumulative probability for that cancer site, which is always less than 1 from the data. The Beta model may be converted into a density by writing $b(t) = C(\gamma t)^{k-1}(1-\beta t)$; where $C = \int(\alpha t)^{k-1}(1-\beta t)dt$, and $\int(\gamma t)^{k-1}(1-\beta t)dt = 1$; $0 \leq t \leq \beta^{-1}$. The factor C might be interpreted as a susceptibility factor, suggesting that a fraction C of the population will contract the site cancer with probability 1 if they live long enough.

REFERENCES

Armitage P and Doll R. 1957. A two-stage theory of carcinogenesis in relation to the age distribution of Human Cancer. Br J Cancer 11(2):161-9

Armitage P and Doll R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis. Br J Cancer 8(1):1-12

Armitage P. 1985. Multistage models of carcinogenesis. Environ Health Perspect 63:195-201

Case RAM. 1966. Tumours of the urinary tract as an occupational disease in several industries. Ann R Coll Surg Engl 39:213-35

Cook PJ, Doll R, and Fellingham SA. 1969. A mathematical model for the age distribution of cancer in man. Int J Cancer 4:93-112

Cox LA. 1995. An exact analysis of the multistage model explaining dose-response concavity. Risk Anal 15(3):359-68

de Rijke JM, Schouten LJ, Hillen HF, *et al.* 2000. Cancer in the very elderly Dutch population. Cancer 89(5):1121-33

Faragher RG and Kipling D. 1998. How might replicative senescence contribute to human ageing? Bioessays 20(12):985-91

Faragher RG. 2000. Cell senescence and human aging: where's the link? Biochem Soc Trans 28(2):221-6

Finkel AM. 1995. A quantitative estimate of the variations in human susceptibility to cancer and its implications for risk management. In: Olin S, Farland W, Park C, *et al.* (eds), Low Dose Extrapolation of Cancer Risks: Issues and Perspectives. ILSI Press, Washington DC, USA

Hart RW and Setlow RB. 1976. DNA repair in late-passage human cells. Mech Ageing Dev 5(1):67-77

Hart RW, D'Ambrosio SM, Ng KJ, *et al.* 1979. Longevity, stability and DNA repair. Mech Ageing Dev 9(3-4):203-23

Hayflick L. 2000. The illusion of cell immortality. Br J Cancer 83(7):841-6

Herrero-Jimenez P, Thilly G, Southam PJ, *et al.* 1998. Mutation, cell kinetics, and subpopulations at risk for colon cancer in the United States. Mutat Res 400(1-2):553-78

Herrero-Jimenez P, Tomita-Mitchell A, Furth EE, *et al.* 2000. Population risk and physiological rate parameters for colon cancer. The union of an explicit model for carcinogenesis with the public health records of the United States. Mutat Res 447(1):73-116

Higami Y and Shimokawa I. 2000. Apoptosis in the aging process. Cell Tissue Res 301(1):125-32

Imaida K, Hasegawa R, Kato T, *et al.* 1997. Clinicopathological analysis on cancers of autopsy cases in a geriatric hospital. Pathol Int 47(5):293-300

Jennings BJ, Ozanne SE, and Hales CN. 2000. Nutrition, oxidative damage, telomere shortening, and cellular senescence: Individual or connected agents of aging? Mol Genet Metab 71(1/2):32-42

Larsen HJ. 1982. Introduction to Probability Theory and Statistical Inference, Third Edition, pp 203-7. John Wiley & Sons, NY, NY, USA

Lechner H, Amort M, Steger MM, *et al.* 1996. Regulation of CD95 (APO-1) expression and the induction of apoptosis in human T cells: changes in old age. Int Arch Allergy Immunol 110(3):238-43

Lee HM, Greeley GHJr, and Englander EW. 2000. Effects of aging on expression of genes involved in regulation of proliferation and apoptosis in the colonic epithelium. Mech Ageing Dev 115(3):139-55

Moolgavkar S, Krewski D, and Schwartz M. 1999. Mechanisms of Carcinogenesis and Biologically Based Models for Estimation and Orediction of Risk. IARC Scientific Publications No 131. International Agency for Research on Cancer, Lyon, France

Moolgavkar SH and Knudsen AG. 1981. Mutation and cancer: a model for human carcinogenesis. J Natl Cancer Inst 66(6):1037-52

Moolgavkar SH. 1978. The multistage theory of carcinogenesis and the age distribution of cancer in man. J Natl Cancer Inst 61(1):49-52

National Vital Statistics Report. Deaths: Final Data for 1997, vol 108, Table 4. (PHS) 99-1120. http://www.cdc.gov/nchs/products/pubs/pubd/nvsr/47-pre/47-pre.htm

Nordling CO. 1953. A new theory on the cancer-inducing mechanism. Br J Cancer 7:68-72

Ogawa T, Kitagawa M, and Hirokawa K. 2000. Age-related changes of human bone marrow: a histometric estimation of proliferative cells, apoptotic cells, T cells, B cells and macrophages. Mech Ageing Dev 117(1-3):57-68

Olkin I, Gleser LJ, and Derman C. 1978. Probability Models and Applications, pp 289-98. MacMillan Publishing, NY, NY, USA

Olofsson P and Kimmel M. 1999. Stochastic models of telomere shortening. Math Biosci 158(1):75-92

Parkin DM, Whelan SL, Ferlay J, *et al.* 1997. Cancer Incidence in Five Continents, vol VII. IARC Scientific Publications No 143. International Agency for Research on Cancer, Lyon, France

Percy C, Van Holten V, and Muir C (eds). 1990. International Classification of Diseases for Oncology, 2nd ed. World Health Organization, Geneva, Switzerland

Pompei F, Polkanov M, and Wilson, R. 2001. Age distribution of cancer in mice: the incidence turnover at old age. Toxicol Indust Health (*in press*)

Potestio M, Caruso C, Gervasi F, *et al.* 1998. Apoptosis and ageing. Mech Ageing Dev 102(2-3):221-37

Ries LAG, Eisner MP, Kosary CL, *et al.* (eds). 2000. SEER Cancer Statistics Review, 1973-1997, National Cancer Institute. Bethesda, MD, USA http://seer.cancer.gov/Publications/CSR1973_1997/

Ries LAG. 2001. Personal communication

Rubelj I and Vondracek Z. 1999. Stochastic mechanism of cellular aging—abrupt telomere shortening as a model for stochastic nature of cellular aging. J Theor Biol 197(4):425-38

Saltzstein SL, Behling CA, and Baergen RN. 1998. Features of cancer in nonagenarians and centenarians. J Am Geriatr Soc 46(8):994-8

Schindowski K, Leutner S, Muller WE, *et al.* 2000. Age-related changes of apoptotic cell death in human lymphocytes. Neurobiol Aging 21(5):661-70

Smith DW. 1996. Cancer mortality at very old ages. Cancer77(7):1367-72
Stanta G, Campagner L, Cavallieri F, *et al.* 1997. Cancer of the oldest old. What we have learned from autopsy studies. Clin Geriatr Med 13(1):55-68
Yancik R and Ries LA. 1994. Cancer in older persons. Cancer (Sup) 74(7):1995-2003