Cancer Informatics

Libertas Academica
FREEDOM TO RESEARCH

ORIGINAL RESEARCH

Weibull-like Model of Cancer Development in Aging

Tengiz Mdzinarishvili and Simon Sherman

Eppley Cancer Institute, University of Nebraska Medical Center, 986805 Nebraska Medical Center, Omaha,
NE 68198-6805. Corresponding author email: ssherm@unmc.edu

Abstract: Mathematical modeling of cancer development is aimed at assessing the risk factors leading to cancer. Aging is a common risk factor for all adult cancers. The risk of getting cancer in aging is presented by a hazard function that can be estimated from the observed incidence rates collected in cancer registries. Recent analyses of the SEER database show that the cancer hazard function initially increases with the age, and then it turns over and falls at the end of the lifetime. Such behavior of the hazard function is poorly modeled by the exponential or compound exponential-linear functions mainly utilized for the modeling. In this work, for mathematical modeling of cancer hazards, we proposed to use the Weibull-like function, derived from the Armitage-Doll multistage concept of carcinogenesis and an assumption that number of clones at age t developed from mutated cells follows the Poisson distribution. This function is characterized by three parameters, two of which (r and λ) are the conventional parameters of the Weibull probability distribution function, and an additional parameter (C_0) that adjusts the model to the observational data. Biological meanings of these parameters are: r—the number of stages in carcinogenesis, λ—an average number of clones developed from the mutated cells during the first year of carcinogenesis, and C_0—a data adjustment parameter that characterizes a fraction of the age-specific population that will get this cancer in their lifetime. To test the validity of the proposed model, the nonlinear regression analysis was performed for the lung cancer (LC) data, collected in the SEER 9 database for white men and women during 1975–2004. Obtained results suggest that: (i) modeling can be improved by the use of another parameter A- the age at the beginning of carcinogenesis; and (ii) in white men and women, the processes of LC carcinogenesis vary by A and C_0, while the corresponding values of r and λ are nearly the same. Overall, the proposed Weibull-like model provides an excellent fit of the estimates of the LC hazard function in aging. It is expected that the Weibull-like model can be applicable to fit estimates of hazard functions of other adult cancers as well.

Keywords: cancer, aging, cancer hazard, Weibull distribution

Cancer Informatics 2010:9 179–188

This article is available from http://www.la-press.com.

Introduction

Mathematical models can help researchers in elucidating the fundamental mechanisms of cancer development and progression. Mathematical models enable a quantitative representation of transformations affecting cell states and cell numbers. One area of cancer modeling is an assessment of risk factors leading to cancer. In adults, aging is a common risk factor associated with cancer development.

The main aim of this work is mathematical modeling of cancer development in aging. Our modeling is based on the commonly accepted multi-stage concept of carcinogenesis and is expected to provide a better understanding of the biological processes underlying cancer development. We utilized observational data on cancer incidence rates collected in the Surveillance Epidemiology and End Results (SEER) database. These incidence rates were used to estimate a cancer hazard function by the log-linear age-period-cohort (LLAPC) model that accounts for age, time period and birth-cohort effects.[1–4]

Until recently, the exponential or compound exponential-linear functions have been widely utilized for mathematical modeling of cancer development in aging.[5–8] According to these mathematical functions, the cancer risk should monotonically increase with aging. However, recent analyses of data collected in the SEER database showed that,[9] after an initial increase, the cancer hazard functions in aging have turnover and even fall at the end of the lifetime.[10] To model such behavior of the cancer hazards, beta-like functions have been utilized.[11–13] It was shown that the observational data can be well-fitted by these functions. The use of beta-like functions, however, still lacks a sound biological background.

In the present work, we proposed to use Weibull-like functions for modeling the cancer hazards in aging. This model is derived from the Armitage-Doll multi-stage concept of carcinogenesis and an assumption that a number of clones at age t that developed from mutated cells follows the Poisson distribution. Besides two conventional parameters (r and λ) of the Weibull probability distribution function (*pdf*), this model has an additional parameter, C_0, for the observational data adjustment. The model provides meaningful biological interpretation of the modeling parameters: r—the number of stages in carcinogenesis, λ—an average number of clones developed from the mutated cells during the first year of carcinogenesis, and C_0—a data adjustment parameter that characterizes a fraction of the age-specific population that will get this cancer in their lifetime.

To test the Weibull-like model of cancer development in aging, we analyzed lung cancer (LC) data collected in the SEER 9 database during 1975–2004 for white men and women. The estimated values of the LC hazard function in aging were obtained from the observed incidence rates. It was found that the quality of modeling has been improved by introducing another parameter, A- the age (in years) at which a process of carcinogenesis starts. Performed modeling suggests that the LC presentation in aging in white men and women are different mainly by the parameters A and C_0, while the parameters r and λ are nearly the same.

Overall, it was shown that the Weibull-like model of cancer presentation in aging provided an excellent fit of the estimates of the LC hazard function in aging. The proposed model explains the observed behavior of the hazard function in aging that initially increases with age, turns over and then falls at the end of the lifetime.

Definitions and Mathematical Statement of the Problem

Survival probability and cancer hazard function

The main concepts of survival analysis include survival probability and hazard function. According to the notation given in textbooks,[14,15] a survival probability is defined as the probability, $S(t)$, that at the time t, a person is alive (in mortality studies) or at the age t a person is free from a given disease (in disease incidence studies). A hazard function, $h(t)$, measures the relative risk of death at the time t, or getting a given disease at a specific age t, compared to the survival probability at the same time/age:

$$h(t) = -\frac{\frac{dS(t)}{dt}}{S(t)} \tag{1}$$

In survival modeling, time t is assumed to follow some statistical distribution with probability density function (*pdf*) – $f(t)$.[15] Let us assume also that $-dS(t)/dt$

can be presented in the form of $C_0 f(t)$ where C_0 is a parameter. If $f(t)$ and C_0 are known, the following differential equation can be written:

$$-\frac{dS(t)}{dt} = C_0 f(t) \qquad (2)$$

with the initial condition

$$S(0) = 1 \qquad (3)$$

This condition means that at time $t = 0$ the person is alive or is free from the disease.

By solving the differential equation (2) with the initial condition (3), one can obtain $S(t)$ by the following formula:

$$S(t) = 1 - \int_0^t C_0 f(z)dz \qquad (4)$$

From (2) and (4), it follows that the formula (1) for a hazard function can be presented as:

$$h(t) = \frac{C_0 f(t)}{1 - \int_0^t C_0 f(z)dz} \qquad (5)$$

If assumption (2) is valid, then, depending on the epidemiological problem under consideration, values of C_0 can vary between 0 and 1. In fact, when $t \to \infty$, formula (4) gives:

$$S(\infty) = 1 - \int_0^\infty C_0 f(z)dz = 1 - C_0 \qquad (6)$$

In mortality studies, $S(\infty) = 0$, when $t \to \infty$. From (6) it also follows that $C_0 = 1$, and therefore the hazard function (5) can be written as:

$$h(t) = \frac{f(t)}{1 - \int_0^t f(z)dz} \qquad (7)$$

In the case of a rare disease, a person's risk of getting a given disease is very small, i.e. survival probability is close to 1:

$$S(t) \cong 1, \quad 0 \le t < \infty \qquad (8)$$

From (8) and (4) it follows:

$$C_0 << 1 \qquad (9)$$

For a given age-specific population, parameter C_0 characterizes a fraction of this population that will be exposed to the disease during their lifetime.

From (4), (5) and (8) it follows that for a rare disease, its hazard function can be presented as:

$$h(t) \cong C_0 f(t) \qquad (10)$$

The aforementioned formulas (4), (5), (7) and (10) for survival probability and hazard function are based on assumption (2). Generally, the validity of this assumption can be tested by the methods of regression analysis using the formula (5), the right side of which presents a regression line with parameters to be determined. In (5), the time t can be considered as a predictor and $h(t)$—as a response variable that can be estimated from the observations. For a rare disease, according to (10), the regression line as a function of t can be approximated by $C_0 f(t)$. Therefore, in the regression analysis performed below we assume that:

$$h(t) = C_0 f(t) \qquad (11)$$

To perform regression analysis, one needs to estimate the values of the response variable $h(t)$. In the absence of time period and cohort effects, the hazard function $h(t)$, can be interpreted as an instantaneous incidence rate.[16] By definition, an incidence rate, $I(t)$, is the number of new cases of cancer incidence over a period of time divided by the person-time at risk.[14] Estimates of values of hazard function, $h^*(t)$, can be obtained from the observed incidence rates that have to be corrected for time period and cohort effects (see below).

Estimation of values of cancer hazards from observations

To estimate the values of a cancer hazard function in aging, the recently proposed method can be utilized.[4,10] This method allows one to correct the observed age-specific incidence rates $I(t)$ for time period and cohort effects. These corrections can be done by the use of the LLAPC model that presents the expectation of the observed incidence rates as a product of the hazard function, $h_c(t)$, the time period effect coefficient, v, and the birth cohort effect coefficient, u. In practice, the observed values of $I(t)$ are presented as $I_{i,j,c}$ (t_i), where t_i is a given age interval, j—a time period interval of observation and c—indicates a given categorical risk factor (for example, gender, race, *etc*.). The procedure allows one: (i) to separate the problem of estimating the time period and birth cohort coefficients from the problem of estimating the unknown hazard function; (ii) to resolve the identifiability problem by an assumption that neighboring cohorts almost equally influence the $I_{i,j,c}(t_i)$ and by anchoring the time period and birth cohort effects to the selected time period and cohort; and (iii) after obtaining the time period and birth cohort coefficients, to estimate values of the hazard function, $h^*_c(t_i)$, and their standard errors, SE_i, in each age interval, t_i.[4,10] Here and below estimates of statistical parameters as well as hazard function values are designated by asterisk (*).

Mathematical statement of the modeling problem

The mathematical problem of modeling of cancer hazard functions can be reduced to the problem of fitting of $h^*_c(t_i)$ by the model curve $h_c(t)$. This implies choosing an appropriate mathematical form of the curve, $(t, C_0f(t))$, which defines a model with the predictor t and the response variable $C_0f(t)$. Thus, this problem is reduced to estimating unknown values of parameters in *pdf*, $f(t)$, and an additional parameter, C_0. These parameters can be estimated by a set of the observed values, $h^*_c(t_i)$, and their standard errors, SE_i, in each age interval, t_i. Estimation of these parameters can be done using the least squares method to solve the following system of the conditional nonlinear equations:

$$h_c^*(t_i) = C_0 f(t_i) \quad i = 1, \dots, n \tag{12}$$

Since each $h^*_c(t_i)$ has its own standard error, SE_i, a system of conditional equations with weights:

$$w_i = \frac{SE^2}{SE_i^2}, \quad i = 1, \dots, n \tag{13}$$

has to be solved. In formula (13), SE is the standard error of the observation with the weight 1, calculated by formula:

$$SE^2 = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{SE_i^2}} \tag{14}$$

In general case, the weighted system, (12)–(14), is solved by a method of least squares, utilizing an iterative technique (see, for example, MATLAB's Statistical Toolbox 7.3, Weighted Nonlinear Regression).[17]

Determination of Mathematical Form of Cancer Hazard in Aging

Weibull model of cancer development in aging

The mathematical model of cancer development is expected to be related to an appropriate biological concept of carcinogenesis. We used the Armitage-Doll multi-stage concept of carcinogenesis and demonstrated that this biological concept mathematically leads to the Weibull-like mathematical form of cancer hazard functions in aging.

According to the Armitage-Doll concept, a normal cell can be transformed into a cancer cell, after the r required mutations occurred within this cell. Below, we applied this concept using the notations and logic described in literature.[18,19]

Let us assume that the process of carcinogenesis begins at the age $t = 0$. Suppose that θ_j is the mutation rate at the j-th gene per year and the probability that a mutation at the j-th gene occurs in a cell prior to the age t is a small number and approximately equals to $\theta_j \cdot t$. Then, a product $\left(\prod_{j=1}^{r}\theta_j\right)\cdot t^r$ estimates the probability that in the given cell the required r mutations occurred prior to time t. The parameter r defines the number of stages in carcinogenesis. The average number of clones that were developed from the mutated cells up to the time t can be presented as:

$$\mu(t) = c_n \cdot \left(\prod_{j=1}^{r}\theta_j\right)\cdot t^r = \lambda \cdot t^r \tag{15}$$

where c_n is the number of cells at risk of mutation. In this formula, the parameter λ is the average number of clones developed from the mutated cells during one year ($t = 1$) after the beginning of carcinogenesis.

Let $N(t)$ be the cumulative number of mutated clones that occurred from age 0 to t and let one follow to the homogeneous Poisson process (HPP).[20] In such a case, according to the definition of HPP, a probability P_0 that the cumulative number of mutated clones is equal to a given number k could be obtained by the formula:

$$P_0\{N(t)=k\} = \frac{(\mu_0(t))^k e^{-\mu_0(t)}}{k!}, \tag{16}$$

where the expected number of mutated clones, $\mu_0(t)$, is proportional to the first order of the age t. In HPP, a cumulative distribution function (*cdf*), $F_0(t)$, of the waiting time to the first occurrence of the mutated clone will be expressed by the formula:

$$F_0(t) = 1 - e^{-\mu_0(t)} \tag{17}$$

In the Armitage-Doll model, however, occurrence of mutation can be considered as a non-homogenous Poisson process (NHPP). In this case, parameter $\mu(t)$ is given by formula (15), where λ and r are constants. It should be noted, that in contrast to HPP, in NHPP the expected number of mutated clones $\mu(t)$ is proportional to the r-th order of the age t.

For the NHPP model, a probability P_0 that the cumulative number of mutated clones equals to a given number k could be obtained by the formula analogous to (16), and *cdf* of the waiting time to the first occurrence of mutated clone can be calculated by formula analogous to (17), where $\mu_0(t)$ is substituted by $\mu(t)$:

$$F(t) = 1 - e^{-\mu(t)} \tag{18}$$

Then, using (15) and (18), the *pdf* of the waiting time t to the first occurrence of cancer of mutated clone, $f(t) = d/dt\ F(t)$ can be presented as a Weibull distribution with the shape parameter r and the combined scale-shape parameter λ:

$$f(t) = \lambda r t^{r-1} \exp(-\lambda t^r) \tag{19}$$

It should be mentioned, however, that in statistical literature, another form of the Weibull *pdf* is usually used.[21,22] This form is:

$$f(t) = \frac{r}{b}\left(\frac{t}{b}\right)^{r-1} \exp\left[-\left(\frac{t}{b}\right)^r\right] \tag{20}$$

where b is the scale parameter, which is related to the parameter λ in (15), by the formula:

$$b = \lambda^{-\frac{1}{r}}. \tag{21}$$

Weibull-like model of age-specific cancer presentation in population

It is known that for most of the cancers, only a tiny fraction, say C_0, of a given age-specific population will (early or later) develop cancer, while the majority of this population will not get this disease in a lifetime.[23] In this case we can assume that a person's overall risk of getting the cancer equals to C_0, and the corresponding hazard function, $h(t)$, can be presented by formula (11). As was shown previously, according to the Armitage-Doll concept, the *pdf* $f(t)$ of cancer occurrence at age t, can be presented by the Weibull distribution in form (19). Therefore, according to (11) and (19), $h(t)$ should have a form of Weibull-like function, which can be presented as a product of Weibull *pdf* and an additional adjustment parameter, C_0:

$$h(t) = C_0 \lambda r t^{r-1} \exp(-\lambda t^r) \tag{22}$$

The parameter r of this function indicates the number of stages in carcinogenesis, the parameter λ is an average number of clones developed from the mutated cells during the first year of carcinogenesis, and parameter C_0 is an adjustment parameter.

Formula (22) states that when age t of cancer occurrence has the Weibull *pdf* (19), then the corresponding hazard function is modeled by the Weibull-like function (22). In contrast to the classical paper,[5] where the exponential function was used for modeling of the cancer hazard in aging, we propose here that the cancer hazard function has a Weibull-like form (22), which has a turnover and falls at old ages.

Validity of our assumption that the $h(t)$ will have a form of Weibull-like function (22) can be tested by methods of regression analysis, considering the age t as a predictor and $h(t)$—as a response variable to be estimated from the observations.

Since in statistical literature the Weibull *pdf* is usually presented by formula (20), the corresponding Weibull-like hazard function can be written as:

$$h(t) = C_0 \frac{r}{b}\left(\frac{t}{b}\right)^{r-1} \exp\left[-\left(\frac{t}{b}\right)^r\right] \quad (23)$$

or

$$h(t) = C_1 \left(\frac{t}{b}\right)^{r-1} \exp\left[-\left(\frac{t}{b}\right)^r\right] \quad (24)$$

where the parameter *r* defines the shape of *h*(*t*), the parameter *b* scales the distribution of *t* along the abscissa axis and the parameter $C_1 = C_0\, r/b$ scales the curve along the vertical axis.

Fitting Weibull-like model by methods of regression analysis

We have shown that cancer hazard functions in aging in a population can be described by the Weibull-like function, presented by the formula (24). The parameters C_1, *b* and *r* can be assessed from bivariate data $(t_i, h_c^*(t_i))$ by the aforementioned least squares method assuming that in this specific case, $h_c^*(t_i)$ can be presented in the following form:

$$h_c^*(t_i) = C_1 \left(\frac{t_i}{b}\right)^{r-1} \exp\left[-\left(\frac{t_i}{b}\right)^r\right] \quad i = 1, \ldots, n \quad (25)$$

Goodness of fitting of the regression line to the observed data can be quantified by the weighted sum-of-squares (*SS*) of the residuals r_i of the system of the corresponding weighted conditional equations:

$$SS = \Sigma_{i=1}^{n} r_i^2 \quad (26)$$

or by the coefficient of determination:

$$R^2 = 1 - \frac{SS}{\Sigma_{i=1}^{n}\left[\sqrt{w_i} h_c^*(t_i) - \frac{1}{n}\Sigma_{i=1}^{n}\sqrt{w_i} h_c^*(t_i)\right]^2} \quad (27)$$

The curve fitting is getting better as R^2 approaches values close to 1.[25]

To compare the quality of fitting of the same dataset by different regression lines, the Akaike's Information Criterion (AIC) can be used.[25] Assuming that the scatter of points around the regression line follows a Gaussian distribution, the AIC is defined by the following equation:

$$AIC = n \ln\left(\frac{SS}{n}\right) + 2K \quad (28)$$

where $K = p + 1$ and *p* is the number of parameters used for curve fitting. When the number of data points *n* are at least two times greater than the number of the assessing parameters, *p*, a second-order (corrected, c) criterion is used:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \quad (29)$$

when the qualities of fitting of the same dataset by different regression lines are compared, the curve fitting is better for the line with the smallest AIC_c.[25]

Weibull-like Model of the Lung Cancer Presentation in Aging

Data preparation and processing

The lung cancer (LC) incidence data, collected in the SEER 9 database during 1975–2004 for white men and women, were used to estimate values of the age-specific hazard function in five year age intervals t_i. Data preparation and filtration were performed by the previously described protocol.[4] LC incidence rates, expressed per 100,000 person per year, were age-adjusted by the direct method to the 2000 United States standard population.[26] Estimates of the age-specific hazard functions $h_c^*(t_i)$ and their standard errors SE_i, anchored to the 2000–2004 time period and to the 1925–1929 birth cohort, were obtained using our recently proposed approach.[4,10]

To perform the curve fitting of bivariate data $(t_i, h_c^*(t_i))$, we modified the aforementioned procedure. In the modified procedure, instead of age *t*, the period of "effective exposure", $t - A$, (where *A* is the age at the beginning of cancer) was utilized. The use of the "effective exposure" period was proposed in literature to improve the quality of curve fitting.[5] In this case, the system of the weighted conditional equations can be defined as:

$$h_c^*(t_i) = C_1\left(\frac{t_i - A}{b}\right)^{r-1} \exp\left[-\left(\frac{t_i - A}{b}\right)^r\right] \quad i = 1, \ldots, n. \quad (30)$$

It should be noted that adding a new "shift" parameter (A) to the three parameters that have been estimated by regression analysis makes this analysis computationally unstable. Our numerical experiments have shown that by introducing an additional parameter, one has dealt with a typical "ill-posed problem", in which small errors in observed data-cause big errors in estimated values. An analogous problem arises in the beta-like modeling of cancer in aging.[27] To avoid such computational instability, we used a method of regularization of solution.[28] For this purpose, by fixing different values of a possible "effective exposure" period (A), we regularized the process of our regression analysis and found the best solution for the other three parameters.

The parameters C_1, b and r can be assessed from bivariate data (t_i, $h_c^*(t_i)$) using the the aforementioned least squares method Specifically, to estimate these parameters we utilized the "*nlinfit*" function from the MATLAB's Statistical Toolbox 7.3. This function allows one to determine the estimates of weighted parameters and perform curve fitting.[24] In the process of the parameter estimation, "*nlinfit*" utilizes an iterative technique that requires one to provide appropriate starting values of parameters to be estimated. Our computational experiments showed that the values 5, 60 and 0.01 can be used as appropriate starting values for the estimates of the parameters, r^*, b^*, and C_1^*, correspondingly.

The output data of the "*nlinfit*" was used as input for two other MATLAB functions, "*nlpredci*" and "*nlparci*". The function "*nlpredci*" returns as output: (i) the error estimates on predictions, *ypred*; and (ii) the half-widths of the 95% prediction intervals for future observations, *delta* (note, 2 × *delta* predicts the observations with the weights of $w = 1$, or the observations with the variance, SE^2). For C_1^*, r^* and b^*, the function "*nlparci*" returns as output the estimates of their errors. To estimate covariance between parameters, the Matlab function "*nlinfit*" provides the covariate matrix, *Sigmaw*. This matrix was used as input for the Matlab function "*nlparci*" to obtain 95% confidence intervals (*CIs*) for parameter estimates.

To evaluate the quality of fitting of the same dataset by different regression lines, we used the Akaike's Information Criterion (AIC) as described previously.

Results and Discussion

The obtained values of the age-specific LC hazard functions and their standard errors for the age groups for which these estimates are statistically distinguishable from zero are presented in Table 1.

Using bivariate data, ($t_i - A$, $h_c^*(t_i)$), we performed curve fitting for several possible "effective exposure" periods. For this purpose, A values were varied from 0 to 30 years with five-year steps. Our calculations showed that for white men the best fitting is achieved when $A = 20$; while for white women it is best when $A = 15$. Table 2 presents the best fitted values of the parameters, C_1^*, r^* and b^*. In addition to the C_1^*, r^* and b^*, the estimates of the parameters, λ^* and C_0^*, are also given in Table 2. These estimates were calculated by the following formulas:

$$\lambda^* = \left(\frac{1}{b^*}\right)^{r^*} \quad (31)$$

$$C_0^* = C_1^* \frac{b^*}{r^*} \quad (32)$$

Table 1. Estimates of the age-specific LC hazard functions, $h^*(t_i)$, and their standard errors, $SE[h^*(t_i)]$, for white men and women.

Age t_i (years)	White men		White women	
	$h^*(t_i)$	$SE[h^*(t_i)]$	$h^*(t_i)$	$SE[h^*(t_i)]$
37.5	1.06	0.04	1.56	0.07
42.5	2.85	0.07	3.78	0.11
47.5	6.95	0.12	8.36	0.18
52.5	13.97	0.19	14.13	0.24
57.5	24.44	0.27	22.09	0.31
62.5	38.38	0.38	29.90	0.37
67.5	53.09	0.48	37.77	0.42
72.5	63.37	0.57	42.26	0.45
77.5	67.09	0.65	41.40	0.47
82.5	57.94	0.77	33.29	0.52
87.5	37.92	0.72	17.20	0.38

Note: In this table $h^*(t_i)$ and $SE[h^*(t_i)]$ are presented per 100,000 person-years.

Table 2. Estimates of model parameters and characteristics of the goodness of the curve fitting.

Parameters and characteristics	White men	White women
	A = 20	A = 15
b^* (95% *CI*)	58.33 (58.08; 58.60)	60.56 (59.94; 61.19)
r^* (95% *CI*)	5.30 (5.25; 5.34)	5.26 (5.09; 5.42)
$\lambda^* \times 10^{10}$ (95% *CI*)	4.40 (5.73; 3.38)	4.27 (1.9; 11.0)
$C_1^* \times 10^5$ (95% *CI*)	179 (177; 181)	114 (110; 119)
$C_0^* \times 10^4$ (95% *CI*)	197 (191; 203)	132 (118; 146)
R^2	0.998	0.935
AIC_C	–289.12	–256.39

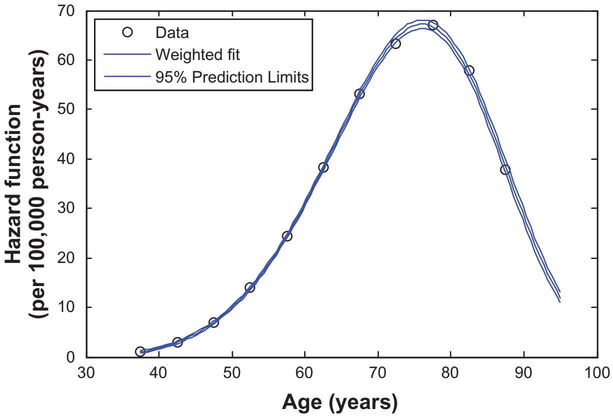

Figure 1. The Weibull-like curve fitted to the estimates (data) of the age-specific LC hazard function for white men. The estimates of the LC hazards are anchored to the 2000–2004 time period and to the 1925–1929 birth cohort.

The corresponding 95% $CI(\lambda^*)$ and 95% $CI(C_0^*)$ were calculated by means of *CI*s of parameters C_1^*, r^* and b^* and formulas (31) and (32).

It should be noted that the *CI*s of the estimates C_1^*, r^* and b^* could be underestimated, because $h_c^*(t_i)$ are not directly observed, but are estimated by observed incidence rates. In practice, when response variable is indirectly observed, to assess *CI*s of parameters of regression, a bootstrap method can be used. Our computational experiments showed, however, that the *CI*s for parameters of the nonlinear regression, obtained by the use of the bootstrap method, were insignificantly different from those obtained by the approach proposed in this paper.

Figure 1 presents the Weibull-like curve that is fitted to the LC observational data for white men, assuming that the effective exposure period starts at the age A = 20. Figure 2 shows an analogous curve fitted to the LC observational data for white women with A = 15. These figures demonstrate that the LC observational data are well-fitted by the Weibull-like functions.

In addition, to verify the goodness of curve fitting, we constructed probability plots for weighted residuals, as described in the statistical textbook.[29] The obtained normal probability plots did not show any significant trends that can be considered as an additional evidence of goodness of fitting.

Based on these fitting data, we hypothesized that the LC carcinogenesis starts earlier in women than in men.

In white men and women, the corresponding estimated values of the parameter r^* are very similar and close to 5. Analogously, the estimated values of the parameter λ that defines the average number of clones, developed from the mutated cells during the first year since the beginning of carcinogenesis, are also similar for both genders. This suggests that in both genders the processes of carcinogenesis in aging are different mainly due to the starting age of carcinogenesis and the fraction of the population exposed to the LC. The starting age of carcinogenesis in white men is about 20 years old and the parameter characterizing the fraction of the age-specific population exposed to the LC is about 197 per 10,000. For white women, the starting age of carcinogenesis is about 15 years old and the parameter characterizing the fraction of the age-specific population exposed to the LC is about 132 per 10,000.

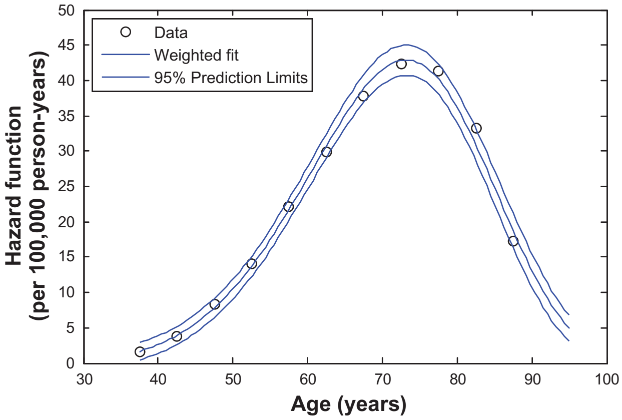

Figure 2. The Weibull-like curve fitted to the estimates (data) of the age-specific LC hazard function for white women. The estimates of the LC hazards are anchored to the 2000–2004 time period and to the 1925–1929 birth cohort.

Conclusion

In this paper, we modeled cancer hazards in aging by the Weibull-like function: $h(t) = C_0 \lambda r t^{r-1} \exp(-\lambda t^r)$. This form of hazard function was derived from the Armitage-Doll multistage concept of carcinogenesis and an assumption that the number of clones developed from the mutated cells follows the Poisson distribution. The proposed modeling function is characterized by three parameters: r and λ are the conventional parameters of the Weibull probability distribution function, and the additional parameter, C_0, adjusts the model to the observational data. These parameters have the following biological meanings: r is the number of stages of carcinogenesis; λ is the average number of clones developed from the mutated cells during the first year since the beginning of carcinogenesis; and C_0 is the data adjustment parameter characterizing a fraction of the age-specific population that will develop the considered type of cancer in their lifetime.

Validity of the Weibull-like model for cancer development in aging was tested by the methods of nonlinear regression analysis using the lung cancer data, collected in the SEER 9 database during 1975–2004 for white men and women. The performed analysis showed that the use of the period of "effective exposure", $t - A$, improved the quality of the modeling. For white men the best quality was obtained when $A = 20$, while for white women the best quality was when $A = 15$. The number of stages of carcinogenesis in white men and women was shown to be similar and close to five, and the average number of clones developed from the mutated cells during the first year since the beginning of carcinogenesis are also similar. The obtained results suggest that in white men and women, the processes of carcinogenesis are different mainly by the starting ages of carcinogenesis and the data adjustment parameters characterizing the corresponding fractions of the age-specific population exposed to the LC.

Overall, we can conclude that the used incidence rate data is consistent with a Weibull model of carcinogenesis that is adjusted for the age of initial cancer exposure. Specifically, the Weibull-like model was shown to fit very well the estimates of the LC cancer hazards in aging that initially increase with the age, turn over and then fall at the end of the lifetime. It is expected that the Weibull-like model can be applicable to other adult cancers as well.

In conclusion, in our work we have studied only age-related correlative factors. These factors should not be directly equated with causation of cancer development. To elucidate causative factors, comprehensive biological models should be further developed.

Acknowledgement

This work was partially supported by 1 R01 CA140940-01 A1 (NIH, SS the PI) grant and an LB506 grant (Nebraska Department of Health, SS the PI). The authors acknowledge Dr. Leo Kinarsky and Mr. Michael X. Gleason for fruitful discussions and help in the preparation of this work for publication. The authors also acknowledge unknown referees whose comments and suggestions helped to improve the manuscript.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Clayton D, Schifflers E. Models for temporal variation in cancer rates. I: age-period and age-cohort models. *Statistics in Medicine*. 1987;6:449–67.
2. Clayton D, Schifflers E. Models for temporal variation in cancer rates. II: age-period-cohort models. *Statistics in Medicine*. 1987;6:469–81.
3. Holford TR. Understanding the effects of age, period, and cohort on incidence and mortality rates. *Statistics in Medicine*. 1991;12:425–57.
4. Mdzinarishvili T, Gleason MX, Sherman S. A novel approach for analysis of the log-linear age-period-cohort model: Application to Lung Cancer Incidence. *Cancer Informatics*. 2009;7:271–80.
5. Cook PJ, Doll R, Fellingham SA. A mathematical model for the age distribution of cancer in man. *Int J Cancer*. 1969;4:93–112.
6. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci U S A*. 2002;99:15095–100.
7. Meza R, Jeon J, Moolgavkar SH, Luebeck EG. Age-specific incidence of cancer: phases, transitions, and biological implications. *Proc Natl Acad Sci U S A*. 2008;105:16284–9.
8. Moolgavkar SH, Meza R, Turim J. Pleural and peritoneal mesotheliomas in SEER: age effects and temporal trends, 1973–2005. *Cancer Causes Control*. 2009;20(6):935–44.
9. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Limited-Use Data (1973–2004), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released Apr 2007, based on the Nov 2006 submission.
10. Mdzinarishvili T, Gleason MX, Sherman S. Estimation of hazard functions in the log-linear age-period-cohort model: application to lung cancer risk associated with geographical area. *Cancer Informatics*. 2010;9:67–78.
11. Mdzinarishvili T, Gleason MX, Sherman S. A generalized beta model for the age distribution of cancers: application to pancreatic and kidney cancer. *Cancer Informatics*. 2009;7:183–97.

12. Harding C, Pompei F, Lee E, Wilson R. Cancer suppression at old age. *Cancer Res*. 2008;68:4465–78.
13. Pompei F, Wilson R. Age distribution of cancer: the incidence turnover at old age. *Hum Ecol Risk Assess*. 2001;7:1619–50.
14. Selvin S. Statistical Analysis of Epidemiologic Data, 3rd Ed. *Oxford University Press*. 2004:1–39.
15. Kleinbaum DG, Klein M. Survival analysis, Second Ed. Springer Science + Business Media, Inc.; 2005:2–43.
16. Jewell NP. Statistics for Epidemiology. Chapman&Hall/CRC; 2004:12–4.
17. http://www.mathworks.com/products/statistics/demos.html?file=/products/demos/shipping/stats/wnlsdemo.html
18. Armitage P, Doll R. The age distribution of cancer and a multistage theory of carcinogenesis, *British Journal of Cancer*. 1954:8:1–12.
19. Klein JP, Andersen PK, Keiding N. Weibull distribution: in Encyclopedia of biostatistics, Armitage P, Colton T editors, 2nd Ed. John Wiley & Sons, Ltd. 2005:6193.
20. *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/apr/section1/apr171.htm
21. Rinne H. The Weibull Distribution: A handbook. *CRC Press Taylor and Francis Group*. 2009:41–2.
22. Krishnamoorthy K. Handbook of statistical distributions with applications. *Charpman&Hall/CRC Taylor and Francis Group*. 2006:263–76.
23. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2006. *CA Cancer J Clin*. 2006 Mar–Apr;56(2):106–30.
24. http://www.mathworks.com/products/statistics/demos.html?file=/products/demos/shipping/stats/cfitdfitdemo.html
25. Motulsky HJ, Christopoulos A. Fitting models to biological data using linear and nonlinear regression: A Practical Guide to Curve Fitting., GraphPad Software Inc., San Diego CA, www.graphpad.com. 2003;32–7 and pp.143–8.
26. Surveillance, Epidemiology, and End Results (SEER) Program. Standard Populations (Millions) for Age-Adjustment [cited 2009 Feb 2]. Available from: http://seer.cancer.gov/stdpopulations/stdpop.singleagesthru99.txt
27. Mdzinarishvili T, Gleason MX, Sherman S. Comment re: Cancer incidence falls for Oldest. *Cancer Res*. 2009;69(1):379.
28. Tikhonov AN, Arsenin VY. Solution of Ill-posed problems. New York; 1977:1–30.
29. Devore JL, Berk KN. Modern Mathematical Statistics with Applications. *Duxbury Press*. 2007:206.