Hindawi Publishing Corporation
Advances in Epidemiology
Volume 2014, Article ID 746769, 8 pages
http://dx.doi.org/10.1155/2014/746769

Research Article

Estimating Cancer Latency Times Using a Weibull Model

Diana L. Nadler and Igor G. Zurbenko

School of Public Health, University at Albany, One University Way, Rensselaer, NY 12144, USA

Correspondence should be addressed to Diana L. Nadler; dnadler@albany.edu

Received 17 April 2014; Revised 10 August 2014; Accepted 11 August 2014; Published 31 August 2014

Academic Editor: Jeanine M. Buchanich

Mathematical models can be useful tools in exploring population disease trends over time and can be used to gain insight into the fundamental mechanisms of cancer development. In this paper, we provide a systematic comparison between the exact and the approximate solutions for estimating the length of time between the biological initiation of cancer and diagnosis through the development of a Weibull-like survival model. A total of 1,608,484 malignant primary cancers were used in the analysis using cancer incidence data obtained from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program. We find that the approximate solution provides a reliable comparison of the latency periods for different types of cancer and has no significant effect on the estimation accuracy, which differs from the exact solution by 0% to 11.3%. Thirty-five of the 44 cancers in this analysis were found to progress silently for 10 years or longer prior to detection representing 89% of the patients in this analysis. The results of this analysis differentiate cancer types that progress undetected over a period of years to identify new opportunities for early detection which increases the likelihood of successful treatment and alleviates the ever-growing cancer burden.

1. Introduction

Cancer is the second leading cause of death in the United States and across the world [1]. It is estimated that 13 million Americans are currently living with cancer and 40.8 percent of men and women can expect to be diagnosed with cancer at some point in their lifetime [2]. In addition to the devastating effects on patients and their families, the economic costs of cancer are enormous, both in terms of direct medical-care resources for its treatment and in the loss of human capital due to early mortality [3]. According to the National Institutes of Health, cancer costs the United States an estimated $263.8 billion in medical costs and lost productivity in 2010 and the cost of cancer care is expected to escalate more rapidly in the near future as more expensive targeted treatments are adopted as standards of care [4].

To estimate the period between the biological initiation of cancer and the medical diagnosis, we utilized the popular two-parameter Weibull distribution as our framework in order to develop the approximate and exact parameter solutions. The Weibull distribution has been used to describe the mechanisms of cancer development in previous research. In contrast to the memoryless exponential distribution which assumes a constant failure rate, the shape of the Weibull distribution is dependent on past events and preserves a memory of prior survival. This provides a simple but powerful way to characterize how the unobserved experience of cancer relates to the observed ones as a function to estimate the time between onset and diagnosis [5].

Using publicly available survival data for breast, lung, pancreatic, and stomach cancers, the approximate solutions are compared with the exact numerical solutions and a comparison of the results are illustrated. A review of the literature available on cancer latency times is provided to validate the conclusions made by the researchers. By providing cancer researchers and policy makers with information about the types of cancers with long latency periods, we identify opportunities for early detection when cancer is most treatable preventing not only mortality but also reducing morbidity and costs. Early detection represents one of the most promising approaches to reducing the growing cancer burden [3].

2. Methods

2.1. Estimating Model Parameters. The Weibull model is widely applied in survival analysis and has been shown to fit data involving the time to appearance of tumors or death in animals subject to carcinogenic insults over time [6]. One of the first contributions to the understanding of cancer incubation periods came about in the 1930s and 1940s when Blum et al. reported that the incubation periods of skin cancer following ultraviolet radiation exposure in mice were log normally distributed [7, 8]. Blum et al. also determined that the "distribution does not vary systematically with dosage of radiation, interval between exposures, intensity of radiation or age" which was one of the first statements about the potential robustness of the incubation period in relation to the dose of exposure [9]. Pike [10] and Peto and Lee [11] also gave a theoretical motivation for the application of the Weibull model to fit data involving the time until the appearance of a tumor or death in animals subject to carcinogenic insults over time [6, 10]. The Weibull model will serve as the underlying framework for this analysis.

Assuming cancer patient survival times follow a Weibull distribution with shape parameter, β, strictly less than 1 and frequently equal to 1/2, we find that the hazard function decreases over time and the distribution has a strong memory of prior survival times. This mathematical property is crucial because it allows us to restore what we cautiously believe to be the length of time between cancer initiation and diagnosis using information known only after diagnosis. The main advantages of the Weibull model extension are its simplicity and ability to promote further research of medical issues through mathematical modeling on a population scale. One well-defined practical application of the study of cancer latency periods is the etiological investigation of cancer which is important in the management of past and future risk assessment for the community and patients [9]. This information can also be advantageous as a means for determining the estimated time of exposure in a patient's life [12].

Using the survival function from the 2-parameter Weibull model $S(t) = e^{-(\lambda t)^{\beta}}$ as a basis, we apply the conditional probability property that $P(A \mid B) = P(A \cap B)/P(B)$ and find that the Weibull random variable $(U+T)$ can be characterized by the conditional survival function:

$$S(u+t \mid u) = \frac{e^{-[\lambda(u+t)]^{\beta}}}{e^{-(\lambda u)^{\beta}}} = e^{-\lambda^{\beta}[(u+t)^{\beta}-u^{\beta}]}; \quad 0 < t. \tag{1}$$

This function represents the probability of surviving beyond time $(u+t)$, given patient survival up to the time of diagnosis. As stated above, the length of the unknown period from malignancy to diagnosis is represented by the lag parameter u; β is the shape parameter; λ is the scale parameter. Each t_i represents the time an individual was observed on the study after diagnosis where study followup ends at death or is right-censored to the study end point.

2.2. Exact Parameter Estimation. Maximum likelihood estimation (MLE) is the most popular method of estimation and is considered to be one of the most versatile and reliable methods [13]. It is one of the most important techniques in statistics and econometrics used for estimation [14]. Researchers can develop new models and estimate these nonstandard statistical model parameters by implementing statistical computing software that utilizes maximum likelihood estimation.

The maximum likelihood estimators of λ, β, and u are the values that maximize the likelihood equation, $L(t, \lambda, \beta, u)$, or equivalently, the logarithm of L, $\log(L(t, \lambda, \beta, u))$. The logarithm of the probability density of the sample (i.e., the log-likelihood) simplifies to

$$\begin{aligned}\log(L(t,\lambda,\beta,u)) &= n\beta\ln(\lambda) + n\ln(\beta) + (\beta-1) \\ &\quad \times \sum_{i=1}^{n}\ln(u+t_i) - \lambda^{\beta}\sum_{i=1}^{n}(u+t_i)^{\beta} + n\lambda^{\beta}u^{\beta}.\end{aligned} \tag{2}$$

To obtain the gradients of the log-likelihood function, we take the partial derivative of the log-likelihood function, $\log(L(t, \lambda, \beta, u))$, with respect to the three parameters λ, β, and u. To maximize the log-likelihood function, we find the parameter values where the partial derivative is equal to zero. The estimating equations with respect to λ, β, and u are

$$\begin{aligned}\frac{\partial\log(L(t,\lambda,\beta,u))}{\partial\lambda} &= n - \lambda^{\beta}\sum_{i=1}^{n}(u+t_i)^{\beta} + n\lambda^{\beta}u^{\beta}, \\ \frac{\partial\log(L(t,\lambda,\beta,u))}{\partial\beta} & \\ &= n\cdot\ln(\lambda) + \frac{n}{\beta} + \sum_{i=1}^{n}\ln(u+t_i) \\ &\quad - \lambda^{\beta}\sum_{i=1}^{n}(u+t_i)^{\beta}\cdot\left\{\ln(\lambda) + \ln\left[\sum_{i=1}^{n}(u+t_i)\right]\Bigg]\right\} \\ &\quad + n\lambda^{\beta}u^{\beta}\left[\ln(\lambda)+\ln(u)\right], \\ \frac{\partial\log(L(t,\lambda,\beta,u))}{\partial u} &= \frac{(\beta-1)}{\sum_{i=1}^{n}(u+t_i)} - \beta\lambda^{\beta}\sum_{i=1}^{n}(u+t_i)^{\beta-1} \\ &\quad + n\beta\lambda^{\beta}u^{\beta-1}.\end{aligned} \tag{3}$$

Taking (3), we define the three-dimensional gradient vector of the log-likelihood function as

$$\begin{aligned}\operatorname{Grad}\log(L(t,\lambda,\beta,u)) &= \left(\frac{\partial\log(L(t,\lambda,\beta,u))}{\partial\lambda}, \frac{\partial\log(L(t,\lambda,\beta,u))}{\partial\beta},\right. \\ &\qquad \left.\frac{\partial\log(L(t,\lambda,\beta,u))}{\partial u}\right).\end{aligned} \tag{4}$$

The gradient of the log-likelihood, presented in (3), provides the formulas to numerically calculate the values of the maximum likelihood parameter estimates using the observed survival times, t_i, which are available through public data. Statistically, we must find the values $\widehat{\lambda}$, $\widehat{\beta}$, and $\widehat{u}$ that maximize the likelihood function based on the given experiment; however, (3) cannot be solved simultaneously because they are a system of nonlinear equations and a closed-form solution does not exist. In such cases, the iterative Newton-Raphson method allows us to calculate the maximum likelihood estimates.

Newton-Raphson is an extremely powerful technique that can be used to find the solution of a system of nonlinear equations. It is one of the most widely used methods for root finding and can be used to maximize multiparameter log-likelihood functions, $l(\theta)$, for θ lying in a parameter space Ω [15]. For such situations, the Newton-Raphson iterative method can be used to find the root of some function and also exhibits rapid quadratic convergence [16]. Another convenient feature of the Newton-Raphson method is the error function that can be estimated using a quadratic approximation.

Many likelihood functions have a unique maximum value at the point $\widehat{\theta}$, which satisfies the equation $\partial l/\partial\theta = 0$. To maximize the likelihood function, the Newton-Raphson method begins with an initial estimation of the location of the root, θ_0, which should be reasonably close to the true root. The method then finds the tangent to the function and extrapolates it to the horizontal intercept of the tangent line to receive θ_1. The point of intersection, θ_1, is taken as the new approximation to the root. This iterative process is continued until the succession of points $\theta_1, \theta_2, \theta_3, \ldots$ converge to $\widehat{\theta}$, when possible. Iteration leads to rapid convergence given that the initial root estimate is located close to the exact root. Mathematically, the kth iteration of the Newton-Raphson algorithm for evaluating the root of some function can be obtained using the formula

$$\theta_{j+1} = \theta_j - H(\theta_j)^{-1} U\left(\theta_j\right) \quad \text{for } j = 1, 2, \ldots. \tag{5}$$

In the formula, the current estimate of the root is θ_j and θ_{j+1} is the extrapolated value of the new root. The vector $U(\theta)$ is the first-derivative vector such that $U(\theta) = \partial l/\partial u$ and $H(\theta)$ is the second derivative, or Hessian matrix, $H(\theta) = \partial^2 l/\partial\theta\partial\theta'$ for $l(\theta)$. By evaluating the function and its derivative at the estimated root, the Newton-Raphson method extrapolates to an improved estimate.

To determine the rate of change, or fractional accuracy, between iterations we can utilize the simple formula

$$e = \frac{\left(x_j - x_{j+1}\right)}{x_{j+1}}. \tag{6}$$

If the fractional accuracy is less than some predetermined value, the algorithm is considered to have converged. A more definitive way to determine convergence is to test the function's value slightly above and below the estimated root to evaluate a change in sign, which could be indicative of a global maximum. A poor initial estimate can give rise to situations where the Newton-Raphson algorithm does not converge or converges slowly [17].

There are free, publicly available software packages such as the R package maxLik that allow researchers to implement maximum likelihood estimation for nonstandard models or the development of new estimators [14]. To optimize the log-likelihood function, researchers must provide the log-likelihood function and a numerical matrix of the starting points for the initial parameter estimates at a minimum. Researchers also have the option to provide the gradient vector and Hessian matrix for faster convergence. If the analytical gradient is not provided by the user, the gradient and Hessian are calculated using the functions numericGradient and numericHessian, which are also included in the maxLik package [14]. The maximization of these functions for simple models works well; however, this may not be the case for complex models so the authors of the maxLik package recommend providing these functions to obtain reliable estimates [14]. To determine the initial starting points for λ_0, β_0, and u_0, we used the approximate parameter estimation method detailed in Section 2.3. The function numericGradient was used to calculate the numeric value of the gradient and evaluate the convergence of the log-likelihood function at successive iterations of the Newton-Raphson algorithm.

We found the estimates provided by the Newton-Raphson algorithm to be unstable due to parameter boundary issues. Since the initial values of the parameters β and λ are close to 0, the Newton-Raphson algorithm was searching for values less than zero which was causing the algorithm to not converge as expected. Although users have the ability to force a parameter value to be equal to some constant when using maxLik, there was no option known to the authors to force the lower boundary of the parameter to be greater than or equal to zero. As a work-around, the authors used the Newton-Raphson algorithm in conjunction with the method of steepest descent (also known as the gradient descent method) to optimize the log-likelihood function. This method searches along the steepest descent direction until the function converges to a stationary point. To validate that the steepest descent method is working correctly, we also verified that the value of the log-likelihood function decreases at each iteration.

2.3. Approximate Parameter Estimation.

Utilizing the methods presented by Nadler and Zurbenko, the approximate parameter value of the latency period can be estimated using simple linear regression methods [5, 18]. Given that the conditional Weibull model is a parametric model, we are able to use likelihood based inference to determine an approximate estimate for the length of time between cancer initiation and diagnosis. Linear regression methods are preferred to estimate the Weibull model parameters because of their computational simplicity and ease of graphical interpretation [19–21]. A linear regression model can be used to estimate the slope, β, because the Weibull model has the key property that the log of the negative log of the estimated survivor function against log time is linear where the regression equation has slope, β, and intercept, $\beta\ln(\lambda)$ [22, 23].

To determine the solution of the MLE of u, denoted as $\widehat{u}$, we find the value of u that maximizes the log-likelihood function proposed in (2), by letting the partial derivative of the log-likelihood with respect to u equal 0:

$$\begin{aligned}&\frac{\partial \log(L(t,\lambda,\beta,u))}{\partial u}\\&\quad= 0 = \sum_{i=1}^{n}\frac{(\beta-1)}{(u+t_i)} - \beta\lambda^{\beta}\left[(u+t_i)^{\beta-1} - u^{\beta-1}\right]\\&\quad= \sum_{i=1}^{n}\frac{(\beta-1)}{(u+t_i)} - \beta\lambda^{\beta}\left[\frac{(u+t_i)^{\beta}}{(u+t_i)} - \frac{u^{\beta}}{u}\right].\end{aligned}\tag{7}$$

To find the approximate MLE solution, we assume that u is large and t is small such that when we evaluate the integral below, we find the following term to be small:

$$\begin{aligned}\int_{t=0}^{\Delta}\frac{\beta-1}{u+t_i}dt &= (\beta-1)\ln(u+t)\Big|_{t=0}^{\Delta}\\&= (\beta-1)\left[\ln(u+\Delta)-\ln(u+0)\right]\\&= (\beta-1)\left[f'(t)\right] \approx (\beta-1)u^{-1}.\end{aligned}\tag{8}$$

We further simplify the approximating equation by substituting the linear regression parameters a and b obtained by regressing the log transform of the Kaplan-Meier product-limit survival estimates on the observed study times for a given sample. The log-survival model takes the form

$$\log\widehat{S}(u+t\mid u) = -\lambda^{\beta}\left[(u+t_i)^{\beta}-u^{\beta}\right] = a + b(u+t_i).\tag{9}$$

Applying these properties to the log-likelihood function, we find the approximate solution of the latency estimate, $\widehat{u}$, by maximizing the equation

$$\begin{aligned}&\frac{\partial \log(L(t,\lambda,\beta,u))}{\partial u}\\&\quad= 0 = \sum_{i=1}^{n}u\cdot\left\{\frac{(\beta-1)}{(u+t_i)} - \beta\lambda^{\beta}\left[\frac{(u+t_i)^{\beta}}{(u+t_i)} - \frac{u^{\beta}}{u}\right]\right\}\\&\quad= \sum_{i=1}^{n}(\beta-1) - \beta\underbrace{\lambda^{\beta}\left[(u+t_i)^{\beta}-u^{\beta}\right]}_{a+b(u+t_i)}\\&\Longrightarrow \frac{-(\beta-1)}{\beta} = a+b(u+t_i).\end{aligned}\tag{10}$$

Letting the partial derivative of the log likelihood with respect to u equal 0, we find the approximate estimate of the length of the latency period at the time of cancer diagnosis, where t_i is equal to 0 as

$$\widehat{u} = \frac{(-(\beta-1)/\beta)-a}{b}.\tag{11}$$

This function represents the time where the log-transformed survival estimate, $S(t)$, regressed on t, equals the correction factor $[-(\beta-1)/\beta]$. By plotting the log negative log Kaplan-Meier estimates against the natural log of time, we can determine the slope of the regression equation, β:

$$\ln\left[-\ln\widehat{S}(u+t\mid u)\right] = \underbrace{-\beta\ln(\lambda)}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}}\ln(u+t_i).\tag{12}$$

To determine model parameters a and b, the log-transformed conditional Kaplan-Meier estimates were regressed on time with intercept, a, and slope, b:

$$\ln\widehat{S}(u+t\mid u) = a + b(u+t_i).\tag{13}$$

3. Analysis

3.1. Data. A retrospective cohort study design was used to study the survival patterns of cancer patients. Cancer incidence and survival data from 1975 through 2008 were obtained from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute. The SEER registries currently collect an estimated 97% of incident cancers within the SEER regions which are representative of the United States population [24]. SEER is a national registry for cancers that is commissioned by the National Cancer Institute which began maintaining records of patients with cancer in 1973 [25]. From this dataset, cancer site, date of diagnosis, summary stage, tumor sequence number, cause of death, and vital status were used in the analysis.

Patients aged 18 and older were included in the analysis if they were diagnosed with their first primary malignant between 1975 and 2008. Patients with multiple malignant tumors over their lifetime were excluded from the cohort, as patients who were known to have died of causes other than cancer were excluded from this analysis as determined by the cause of death variable available in the SEER dataset. Invasive malignant neoplasms, as considered by SEER where the stage is either localized, regional, or distant, were included in this analysis and in situ cases were excluded. The types of cancer chosen for this analysis were restricted to those with high mortality rates and limited availability of effective treatment options allowing the disease to follow its natural course minimizing potential biases. By selecting cancers with high mortality rates, we maximize the amount of information known to the researcher allowing more precise estimates. Overall, 44 cancer types were selected and analyzed with a total sample size of 1,608,484 patients.

3.2. Results. We iteratively solve for the maximum likelihood estimates, $\widehat{\lambda}$, $\widehat{\beta}$, and $\widehat{u}$ using the Newton-Raphson algorithm. We may use any reasonable estimate of the parameters as our initial starting point, λ_0, β_0, and u_0. We chose to use the estimates obtained from the approximate solution which can be solved using the linear regression models presented in (12) and (13). By regressing the log Kaplan-Meier estimates on the observed study times for a given sample, we find the intercept and slope of the linear regression line to find the estimates for a and b. We also regress the log negative log Kaplan-Meier survival estimates on the natural log of time to determine the slope and shape parameters of the Weibull model, β_0 and λ_0, respectively. To determine the initial starting point for the latency parameter, u_0, we use the formula presented in (11).

Using the starting values of the parameters, λ_0, β_0, and u_0, the Newton-Raphson algorithm was used to determine the next starting point, $\lambda_{\mathrm{NR}(1)}$, $\beta_{\mathrm{NR}(1)}$, and $u_{\mathrm{NR}(1)}$. We find the Newton-Raphson algorithm to be unstable when the starting values of the parameters λ_0 and β_0 are close to zero. In order to receive a better approximation of the parameters, we used a combination of the Newton-Raphson algorithm and the gradient method. The gradient is a vector that is directed towards the maximum value of the likelihood equation. To determine the maximum likelihood estimates, we use the values $\lambda_{\mathrm{NR}(1)}$, $\beta_{\mathrm{NR}(1)}$, and $u_{\mathrm{NR}(1)}$ as the starting point for the gradient method. At each successive iteration, the direction of the gradient is calculated and followed to the peak of the likelihood function expecting that at each iteration, the values of the gradient vector decrease. By definition, the maximum likelihood values occur at the point at which the vector is zero.

The accuracy of the estimates provided can be improved only by increasing the study sample size to decrease the variability of the estimate, $\sigma/\sqrt{n}$; therefore, the preciseness of our estimates are constrained by the sample size for publicly available data. We may find that after several iterations of the gradient method we must stop iterating because the parameter estimates are the most precise we are able to obtain out of the sample that is available to us. The approximate and exact solutions for breast, lung, pancreatic, and stomach cancers are provided in Table 1.

We find that the exact solutions are very close to the approximate solutions which differ by approximately 0% to 11.3%. The key feature of the approximate solution is that it provides a reliable solution that is obtainable using simple linear regression which is the most widely used of all statistical techniques. It is important to mention that it is always possible to obtain the exact solution when considered necessary; however, determining the exact solution is very complex, time consuming, and not easily automatable.

In Table 2, we determine the total sample size, 5-year survival rate, approximate time from cancer initiation to diagnosis, and the median age at cancer initiation by cancer site. The approximate estimates include all invasive cancer stages combined to determine the average length of time from initiation to diagnosis for the entire population. These estimates depend on the totality of cancer treatment and are dependent on the health system infrastructure, availability of resources, and overall health of the population; therefore, the methods presented can be used with other data sources but the results may differ due to these factors. The information presented in Table 2 can be used to determine types of cancer where early detection is most impactful; those with a high case volume, poor survival rate, and the growth of the cancer go undetected for an extended period of time.

Thirty-five of the 44 cancers presented progress for 10 years *or more* prior to detection representing 89% of the 1,608,484 total malignancies in this analysis. These results are extremely encouraging as they identify a multitude of opportunities for new research on early detection and preventative screening. It is also useful to reference the median age at initiation to determine when the cancer initiated to advise effective screening guidelines and practices which is valuable information for public health professionals.

Table 1: Comparison of approximate and exact latency estimates for selecting cancers by site.

Cancer site	Exact latency period (years)	Approximate latency period (years)	% difference
Breast	17.2	16.3	5.2%
Lung	13.6	13.6	0.0%
Pancreas	9.5	8.43	11.3%
Stomach	22.9	22.3	2.6%

This information is intended to raise questions rather than to answer them. It is important to understand at the outset that the availability of scientific literature assessing the minimum latency period for specific types of cancer is scarce [26]. To evaluate the validity of our model estimates, we review literature involving the biological and statistical estimation of the length of cancer latency periods.

A recent journal article identified an alarming national surge in the number of cases of oropharyngeal cancers in people under the age of 45 and the human papillomavirus, or HPV, may be what is causing the increase [27]. Historically, oropharyngeal cancers have been predominant in elderly people with a prolonged history of heavy smoking or drinking with a typical latency period of 25 years [28, 29]. Gayar et al. state that the predominance of oropharyngeal cancers in patients under 45 "suggests either nonsexual modes of HPV transfer at a younger age or a shortened latency period between infection and development of cancer" [27]. The results in Table 2 support this finding and indicate a 12.3 year latency period for oropharyngeal cancer and 16.9 years for other oral cavity and pharynx cancers suggesting that a shortened latency period may be partly blamed for this sudden epidemic.

Ovarian cancer, which is the fifth leading cause of cancer death in women aged 35 to 74 years, is usually found at an advanced stage causing the survival rate to be lower than other types of cancer that are easier to detect at an early stage [30]. Ovarian cancer has a reasonably long latency period between initiation and manifestation of established disease which is exacerbated by the late detection of the disease which has been estimated to be approximately 30–40 years for advanced ovarian cancer [31]. Another source cites that the risk of ovarian cancer due to acute radiation exposure at age 10 is approximately three times greater than exposure at age 50 and the latency period for solid tumors can range from 5 to 40 years, with a period of expression longer than 50 years for some cancers [32] which coincides with the approximated latency estimate of 44.1 years for ovarian cancer found in Table 2.

Table 2: Approximate latency times from cancer initiation to diagnosis by cancer site.

Cancer site	Sample size	5-year survival rate	Years from onset to diagnosis	Median age at cancer onset
Acute lymphocytic leukemia	3,701	21.5%	35.7	8.3
Acute monocytic leukemia	1,118	8.8%	15.7	47.3
Acute myeloid leukemia	17,733	12.3%	25.7	39.3
Aleukemic, subleukemic, and NOS	1,785	15.5%	19.3	52.7
Ascending colon	30,038	46.2%	56.8	16.2
Brain	36,828	9.9%	21.9	36.1
Breast	378,477	54.3%	16.3	43.7
Cecum	46,552	36.7%	52.4	20.6
Chronic lymphocytic leukemia	24,466	15.9%	2.2	67.8
Chronic myeloid leukemia	10,498	9.6%	5.1	58.9
Descending colon	13,634	42.4%	52.4	16.6
Esophagus	26,504	6.0%	9.4	56.6
Floor of mouth	5,260	31.5%	21.9	40.1
Gallbladder	8,105	9.6%	25.2	46.8
Gum and other mouth	9,834	37.5%	28.7	34.3
Hypopharynx	5,241	4.8%	9.6	53.4
Kidney and renal pelvis	56,093	33.2%	48.2	14.8
Large intestine, NOS	9,225	19.0%	37.9	36.1
Larynx	22,545	43.1%	35.4	27.6
Liver	22,316	6.0%	10.8	53.2
Lung and bronchus	358,750	6.4%	13.6	53.4
Myeloma	33,252	3.8%	3.6	65.4
NHL-nodal	70,558	27.5%	26.5	37.5
Nasopharynx	4,435	32.4%	25.2	29.8
Nose, nasal cavity, and middle ear	4,062	30.6%	23.0	40.0
Oropharynx	1,763	18.6%	12.3	48.7
Other biliary	8,811	7.4%	16.1	54.9
Other digestive organs	2,145	7.3%	6.6	63.4
Other myeloid/monocytic leukemia	1,424	13.1%	10.5	61.5
Other oral cavity and pharynx	1,722	14.9%	16.9	46.1
Ovary	47,721	25.8%	44.1	17.9
Pancreas	65,835	1.8%	8.4	60.6
Peritoneum, omentum, and mesentery	2,622	11.5%	9.2	56.8
Pleura	5,153	2.5%	4.5	65.5
Rectosigmoid junction	28,603	35.8%	36.6	31.4
Rectum	60,514	34.3%	29.8	37.2
Retroperitoneum	2,433	12.0%	23.1	37.9
Sigmoid colon	69,135	44.3%	52.1	16.9
Small intestine	7,957	25.9%	26.0	38.0
Splenic flexure	7,259	31.7%	42.4	27.6
Stomach	54,521	11.9%	22.3	46.7
Tongue	14,102	33.8%	26.7	33.3
Tonsil	7,429	27.5%	18.9	39.1
Transverse colon	18,325	37.8%	57.0	15.0

A biological study published in *Nature* collected genetic materials from 7 patients who died of end-stage pancreatic cancer and determined the timing of carcinogenesis. Researchers found that it took 11.7 years, on average, for a mature pancreatic tumor to form after the appearance of the first cancer-related mutation in a pancreatic cell. Another 6.8 years passed, on average, before the primary tumor sent out a metastatic lesion to another organ. From that point, the patient died in 2.7 years, on average. In total, more than 20 years elapsed between the appearance of the first mutated pancreatic cell and death [33–36]. The estimate obtained using the Weibull model extension indicates that

8.4 years passed on average from the time of cancer initiation to diagnosis for patients with all-stage pancreatic cancers combined.

Colorectal cancer is another commonly diagnosed cancer in both men and women and is the third leading cause of cancer death. The American Cancer Society states that the majority of colorectal cancers and deaths could be prevented by applying existing knowledge about cancer prevention and increasing the use of established screening tests [37]. There has been sufficient evidence established to conclude that smoking causes colorectal cancer which appears to be stronger for rectal cancer than for colon cancer with a particularly long latency period [37, 38]. The Health Professionals and Nurses' Health study showed an association of at least three to four decades between tobacco exposure and colorectal cancer diagnosis [39, 40]. It is believed that previous research had difficulty identifying this association due to the long latency period and also because of the association due to the extensive time lag between smoking and the occurrence of an adenocarcinoma [41]. The latency estimates provided in Table 2 range from 29.8 years for rectal cancer to 57 years for traverse colon cancer which complement existing research in the field.

4. Conclusions

The approximate solution provides a reliable assessment of the length of time from what is cautiously believed to be the biological initiation of cancer until diagnosis where cancer's growth is often silent and undetected. The approximate solution provides a very good comparison between the different types of cancer and the small difference between the exact estimates do not affect key scientific inferences. When required, the exact solution is always obtainable through the methods described in this paper. This prediction method is recommended for cancer types with a large number of patients and a high mortality, so the true shape of the survival distribution is known and reduces the potential for bias. For example, melanoma cancer was excluded from our estimates because we believe it to be an unreliable estimate because advanced detection methods have drastically changed the shape of the true survival experience due to longer life expectancies and fewer deaths attributed to skin cancers.

The results also have implications for researchers and policy makers when appropriating resources. The approximate solution can also provide a scientific basis for screening guidelines and diagnostic tests. It is important to remember that cancers may be silent and not associated with any symptoms until late-stage cancer develops which is why screening and early detection are so important. Patients often report nonspecific symptoms, delaying the diagnosis of the root cause of cancer which greatly decreases the chance of survival. The results indicate that malignant cells are present in the body for decades but may be undetectable using currently available technologies which identify a number of opportunities for new research on early detection and preventative screening. The cost of cancer care in the United States is substantial and is expected to increase due to population changes alone, and early detection provides one of the most promising approaches to reduce the growing cancer burden [3, 4].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] National Center for Chronic Disease Prevention and Health Promotion, "Leading Causes of Death," January 2013, http://www.cdc.gov/nchs/fastats/lcod.htm.

[2] National Cancer Institute, *SEER Stat Fact Sheets: All Cancer Sites*, 2013.

[3] R. Etzioni, N. Urban, S. Ramsey et al., "The case for early detection," *Nature Reviews Cancer*, vol. 3, no. 4, pp. 243–252, 2003.

[4] A. B. Mariotto, K. Robin Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown, "Projections of the cost of cancer care in the United States: 2010–2020," *Journal of the National Cancer Institute*, vol. 103, no. 2, pp. 117–128, 2011.

[5] D. L. Nadler and I. G. Zurbenko, "Developing a Weibull model extension to estimate cancer latency," *ISRN Epidemiology*, vol. 2013, Article ID 750857, 6 pages, 2013.

[6] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer Science+Business Media, New York, NY, USA, 2003.

[7] H. F. Blum, H. G. Grady, and J. S. Kirby-Smith, "Limits of accuracy in experimental carcinogenesis as exemplified by tumor induction with ultraviolet radiation," *Journal of the National Cancer Institute*, vol. 3, no. 1, pp. 83–89, 1942.

[8] H. F. Blum, H. G. Grady, and J. S. Kirby-Smith, "Relationships between dosage and rate of tumor induction by ultraviolet radiation," *Journal of the National Cancer Institute*, vol. 3, no. 1, pp. 91–97, 1942.

[9] H. K. Armenian, "Incubation periods of cancer: old and new," *Journal of Chronic Diseases*, vol. 40, supplement 2, pp. 9S–15S, 1987.

[10] M. C. Pike, "A method of analysis of a certain class of experiments in carcinogenesis," *Biometrics*, vol. 22, no. 1, pp. 142–161, 1966.

[11] R. Peto and P. N. Lee, "Weibull distributions for continuous-carcinogenesis experiments," *Biometrics*, vol. 29, no. 3, pp. 457–470, 1973.

[12] S. Cobb, M. Miller, and N. Wald, "On the estimation of the incubation period in malignant disease. The brief exposure case, leukemia," *Journal of Chronic Diseases*, vol. 9, no. 4, pp. 385–393, 1959.

[13] H. Rinne, *The Weibull Distribution: A Handbook*, CRC Press, Boca Raton, Fla, USA, 2010.

[14] A. Henningsen and O. Toomet, "maxLik: a package for maximum likelihood estimation in R," *Computational Statistics*, vol. 26, no. 3, pp. 443–458, 2011.

[15] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, Hoboken, NJ, USA, 2003.

[16] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematic Statistics*, Pearson Prentice Hall, Upper Saddle River, NJ, USA, 2005.

[17] S. L. Morgan, *Acid-Base Equilibria: Solving the Cubic Equation by the Newton-Raphson Method*, 2000.

[18] D. L. Nadler and I. G. Zurbenko, "Model prediction of the length of cancer prior to diagnosis with application to cancer registry data," in *JSM Proceedings, Biometrics Section*, pp. 5404–5414, American Statistical Association, Miami Beach, Fla, USA, 2011.

[19] D. N. P. Murthy, M. Bulmer, and J. A. Eccleston, "Weibull model selection for reliability modelling," *Reliability Engineering and System Safety*, vol. 86, no. 3, pp. 257–267, 2004.

[20] L. F. Zhang, M. Xie, and L. C. Tang, "A study of two estimation approaches for parameters of Weibull distribution based on WPP," *Reliability Engineering and System Safety*, vol. 92, no. 3, pp. 360–368, 2007.

[21] L. F. Zhang, M. Xie, and L. C. Tang, "Bias correction for the least squares estimator of Weibull shape parameter with complete and censored data," *Reliability Engineering and System Safety*, vol. 91, no. 8, pp. 930–939, 2006.

[22] D. G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*, Spring Science+Business Media, New York, NY, USA, 2005.

[23] G. Rodriguez, *Lecture Notes on Generalized Linear Models*, 2007.

[24] K. M. Fairfield, K. Murray, F. L. Lucas et al., "Completion of adjuvant chemotherapy and use of health services for older women with epithelial ovarian cancer," *Journal of Clinical Oncology*, vol. 29, no. 29, pp. 3921–3926, 2011.

[25] Surveillance; Epidemiology; and End Results (SEER) Program, *Research Data (1973–2008)*, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, 2011.

[26] J. Howard, *Minimum Latency & Types or Categories of Cancer*, World Trade Center Health Program, 2013.

[27] O. H. Gayar, J. J. Ruterbusch, M. Elshaikh et al., "Oropharyngeal carcinoma in young adults: an alarming national trend," *Otolaryngology—Head and Neck Surgery*, vol. 150, no. 4, pp. 594–601, 2014.

[28] A. C. Nichols, D. A. Palma, S. S. Dhaliwal et al., "The epidemic of human papillomavirus and oropharyngeal cancer in a Canadian population," *Current Oncology*, vol. 20, no. 4, pp. 212–219, 2013.

[29] L. J. Kleinsmith, *Principles of Cancer Biology*, Pearson Benjamin Cummings, San Francisco, Calif, USA, 2005.

[30] J. M. Torpy, A. E. Burke, and R. M. Golub, "Ovarian cancer," *Journal of the American Medical Association*, vol. 305, no. 23, article 2484, 2011.

[31] D. M. Purdie, C. J. Bain, V. Siskind, P. M. Webb, and A. C. Green, "Ovulation and risk of epithelial ovarian cancer," *International Journal of Cancer*, vol. 104, no. 2, pp. 228–232, 2003.

[32] J. T. Bushberg, J. Anthony Seibert, E. M. Leidholdt, and J. M. Boone, *The Essential Physics of Medical Imaging*, Lippincott Williams & Wilkins, Philadelphia, Pa, USA, 2012.

[33] American Cancer Society, Cancer Facts & Figures, 2011.

[34] S. Yachida, S. Jones, I. Bozic et al., "Distant metastasis occurs late during the genetic evolution of pancreatic cancer," *Nature*, vol. 467, no. 7319, pp. 1114–1117, 2010.

[35] *Pancreatic Cancer Grows Over 20 Years*, 2011.

[36] R. Parker, *Pancreatic Cancer Develops for 20 Years Before Killing*, 2010.

[37] American Cancer Society, *Colorectal Cancer Facts & Figures 2011–2013*, American Cancer Society, Atlanta, Ga, USA, 2011.

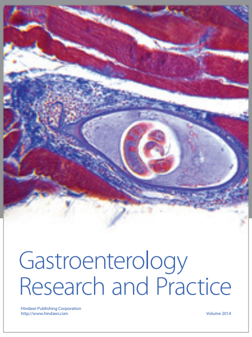
[38] P. S. Liang, T. Y. Chen, and E. Giovannucci, "Cigarette smoking and colorectal cancer incidence and mortality: systematic review and meta-analysis," *International Journal of Cancer*, vol. 124, no. 10, pp. 2406–2415, 2009.

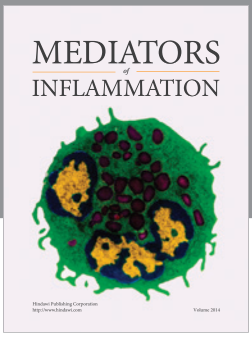
[39] E. Giovannucci, "An updated review of the epidemiological evidence that cigarette smoking increases risk of colorectal cancer," *Cancer Epidemiology Biomarkers and Prevention*, vol. 10, no. 7, pp. 725–731, 2001.

[40] E. Giovannucci and M. E. Marti, "Tobacco, colorectal cancer, and adenomas: a review of the evidence," *Journal of the National Cancer Institute*, vol. 88, no. 23, pp. 1717–1730, 1996.

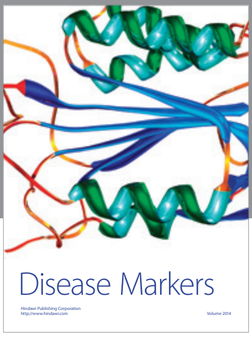
[41] M. Lüchtenborg, K. K. L. White, L. Wilkens, L. N. Kolonel, and L. Le Marchand, "Smoking and colorectal cancer: different effects by type of cigarettes?" *Cancer Epidemiology Biomarkers and Prevention*, vol. 16, no. 7, pp. 1341–1347, 2007.

The Scientific World Journal

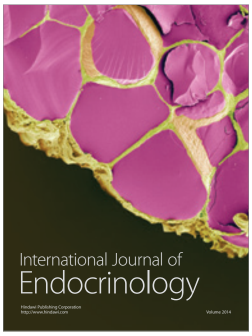

Gastroenterology Research and Practice


MEDIATORS of INFLAMMATION

Journal of Diabetes Research


Disease Markers

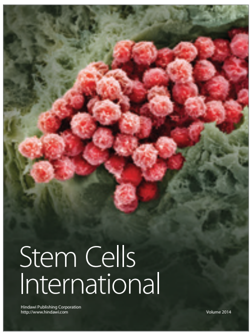
Journal of Immunology Research

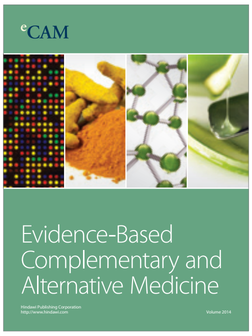

International Journal of Endocrinology

PPAR Research

BioMed Research International

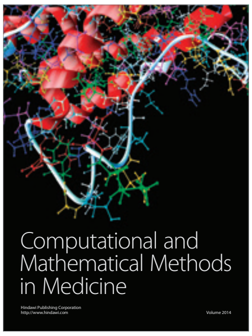
Journal of Ophthalmology

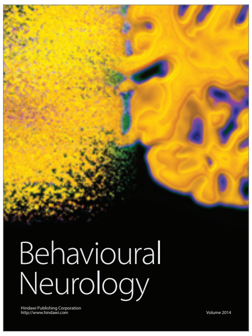

Stem Cells International

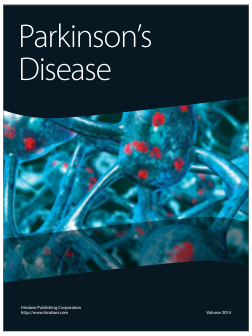

eCAM
Evidence-Based Complementary and Alternative Medicine

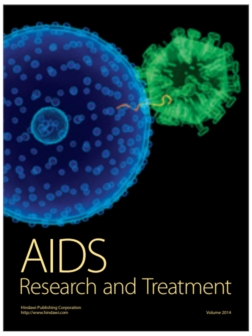
Journal of Obesity

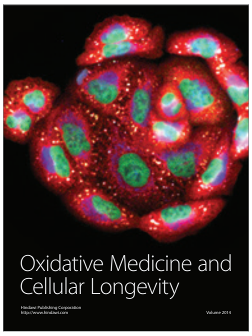
Journal of Oncology


Computational and Mathematical Methods in Medicine


Behavioural Neurology


Parkinson's Disease


AIDS Research and Treatment


Oxidative Medicine and Cellular Longevity