Title
A theory of the cancer age-specific incidence data based on extreme value distributions

Permalink
https://escholarship.org/uc/item/3xq4j5qg

Journal
AIP Advances, 2(1)

ISSN
21583226

Authors
Soto-Ortiz, Luis
Brody, James P

Publication Date
2012

DOI
10.1063/1.3699050

A theory of the cancer age-specific incidence data based on extreme value distributions

Luis Soto-Ortiz and James P. Brody[a]
Department of Biomedical Engineering, Center for Complex Biological Systems, University of California–Irvine, Irvine, California 92697-2715, USA

(Received 31 August 2011; accepted 28 November 2011; published online 22 March 2012)

The incidence of cancers varies with age, if normalized this is called the age-specific incidence. A mathematical model that describes this variation should provide a better understanding of how cancers develop. We suggest that the age-specific incidence should follow an extreme value distribution, based on three widely accepted assumptions: (1) a tumor develops from a single cell, (2) many potential tumor progenitor cells exist in a tissue, and (3) cancer is diagnosed when the first of these many potential tumor cells develops into a tumor. We tested this by comparing the predicted distribution to the age-specific incidence data for colon and prostate carcinomas collected by the Surveillance, Epidemiology and End Results network of 17 cancer registries. We found that colon carcinoma age-specific incidence data is consistent with an extreme value distribution, while prostate carcinomas age-specific incidence data generally follows the distribution. This model indicates that both colon and prostate carcinomas only occur in a subset of the population (22% for prostate and 13.5% for colon.) Because of their very general nature, extreme value distributions might be applicable to understanding other chronic human diseases. [http://dx.doi.org/10.1063/1.3699050]

I. INTRODUCTION

Cancer is a genetic disease.[1,2] A tumor develops after a cell accumulates a number of alterations to its DNA.[3] Most cancers develop from a single cell.[4] The exact process by which a tumor develops is not known, but many potential tumor progenitor cells exist in a tissue. These potential progenitor cells accumulate DNA alterations that transform the normal cell into a tumor cell.[5] When the *first* of these many potential progenitor cells acquires a sufficient set of alterations, a cancer develops.

The age-specific incidence for cancer can be easily measured.[6] The number of diagnosed cancers, as a function of age, is compiled and then divided by the total population at risk, as a function of age. The age-specific incidence is traditionally reported as a hazard function, rather than a probability distribution. The denominator (population at risk) is equal to the total population minus the number who have already developed the cancer. The difference between the hazard function and the probability distribution is insignificant when dealing with carcinoma in the general population.

In 1954, at a time when little was known about the cellular nature of cancer, Armitage and Doll published a theory of the age-specific incidence.[7,8] They noted that the specific incidence of many common cancers increase as a power law as a function of age. They suggested that if cancers were caused by a series of mutations, then the age-specific incidence, when plotted on a log-log graph, should have a slopes equal to one less than the number of mutations required to transform the cell into a tumor cell. This model became widely accepted and appears in textbooks today[3,9]

Further work has extended the Armitage-Doll approach. Models have evolved to incorporate clonal expansion.[10] The culmination of this line of research is the two step with clonal expansion

[a] Author to whom correspondence should be addressed. Electronic mail: jpbrody@uci.edu

model.[11,12] Other approaches to the problem also exist. These include using a Weibull distribution for lung cancer,[13] analyzing the age-specific acceleration of cancers,[14,15] modifying the Armitage-Doll equation directly with a damping term,[16–18] and using a multistage model with age-dependent behavior to estimate the number of mutations required to develop breast cancer.[19] Most of these models implicitly assume that a single cell per tissue has the potential to develop into a tumor.

The Armitage-Doll model was developed before it was known that a tumor develops from a single cell within a tissue. Thus, this model and those that evolved from it misses an important point: many many potential progenitor cells exist, but cancer is diagnosed when the *first* tumor forms. Extreme value theory addresses this issue.

The age-specific incidence for prostate and colon carcinoma are very different. The incidence of prostate carcinoma is essentially zero before the age of 30, but rises to 1000 per 100,000 by the age of 73, then decreases. In contrast, the incidence of colon carcinoma rises much more slowly. It is small, but measurable at age 20. It increases exponentially with age to the highest measured age, 84. Most other solid tumors have an age-specific incidence very similar to colon carcinoma. The goal of this work is to develop and test a theory of the age-specific incidence data for colon and prostate carcinomas.

A. Development of the theory

The theory is based on three well established assumptions:

1. Colon and prostate tumors are clonal; they originate from a single cell.
2. Many potential tumor progenitor cells exist in every tissue.
3. A cancer is diagnosed when the first of these many potential tumors progenitor cells acquires a sufficient set of mutations to transform into a tumor.

Based on these assumptions, we propose that the probability of developing a particular tumor, as a function of time, should follow an extreme value distribution. In contrast to the Armitage-Doll model, this approach does not in any way rely upon the details of how a cell acquires its mutations or the number of mutations required to transform the normal cell into a tumor cell.

An extreme value distribution characterizes the maximum (or minimum) of a parent distribution. If a series of values are randomly drawn from a population (characterized by the parent distribution), the minimum follows an extreme value distribution. In this case, each potential progenitor cell follows the unknown parent distribution. The exact parent distribution would depend upon the number of mutations required, clonal expansion, and other physiological details of how a normal cell transforms into a tumor cell, which is not known. However the particular form of the parent distribution is irrelevant to its extreme value distribution.

Specifically, the Fisher-Tippett theorem[20] guarantees that the distribution of the first occurrence will follow one of three distributions in the limit where the number of samples from the parent distribution is large. The extreme value distribution will be either the Gumbel, Frechet, or Weibull distribution.

The specific form of the extreme value distribution is determined by some general properties of the parent distribution. If the parent distribution has unbounded tails and finite moments, the Gumbel distribution applies. If the parent distribution has some moments that do not exist, then the Frechert distribution applies. If the parent distribution is both bounded and the moments exist, then the Weibull distribution applies. In this case, the Weibull distribution is the appropriate choice of the three, because the age-specific incidence data is bounded on one side (no one can get cancer before conception) and all moments exist.

The Weibull distribution can be written as,

$$p(t) = A\left(\frac{k}{\lambda}\right)\left(\frac{t-\tau}{\lambda}\right)^{k-1}\exp^{-(\frac{t-\tau}{\lambda})^{k}}, \tag{1}$$

for $t \geq \tau$ and $p(t) = 0$ for $t < \tau$. The Weibull distribution has four parameters: A is a normalization factor, τ is a time shift, k is known as the shape parameter, and λ is called the scale parameter. The shape and scale parameters must be positive numbers.

We tested this model in three ways. First, we performed numerical simulations to confirm that the Weibull distribution is the appropriate extreme value distribution to use. Second, we compared weibull to age-specific incidence data for prostate and, third, for colon carcinomas.

II. METHODS

For the numerical simulations, we selected 1000 random numbers from a normal distribution with a mean value of 1000 and a standard deviation of 100. We recorded the minimum of these 1000 numbers. We repeated this process 20000 times to generate the simulated data set. The choice of the mean and standard deviation are arbitrary.

The colon and prostate carcinoma data sets were acquired from the Surveillance, Epidemiology, and End Results (SEER) network of cancer registries. The SEER program of the National Cancer Institute (NCI) is considered the gold-standard for data quality for cancer registries. It collects information on cancer cases from seventeen different geographic areas of the United States encompassing about 26% of the population of the United States.

For colon carcinoma data we selected from the SEER database[21] all women who were diagnosed with colon carcinoma (excluding sarcomas and other forms of cancer) in 2008, the most recent year available. We only used women because women have a slightly different incidence than men, and there are more women at older ages, increasing the statistical power of the data.

For the prostate carcinoma data we selected all men who were diagnosed with prostate carcinoma in 2008. This excludes other forms of prostate cancer, like prostate sarcomas that probably originate through a different process.

To test whether the model fit the data, we performed a least squares fit to determine the parameters that minimized the chi-squared function,

$$\chi^2 = \frac{1}{k}\sum_{t=1}^{k}(p(t; A, k, \lambda, \tau) - o_t)^2, \tag{2}$$

which is the difference between the model's predictions, $p(t;\, A,\, k,\, \lambda\tau)$, and the observed values, o_t, summed over all relevant time points. The relevant time points included any with at least 10 observed counts ($o_t > 9$). The minimization used the generalized reduced gradient algorithm of minimization.[22] We used multiple starting points for the parameters to ensure it does not converge on a relative minimum.

III. RESULTS

We first tested this hypothesis by comparing simulated data to the Weibull distribution. The data consisted of 20,000 minima, as shown in Figure 1. We fit this data to the Weibull equation with parameters $A = 20223$, $k = 23.1$, $\tau = 0$ and $\lambda = 691.9$. This fit had a $\chi^2 = 187$ with 147 degrees of freedom, indicating excellent agreement.

With colon carcinoma, we tested two hypothesis: (1) the total population is susceptible to colon carcinoma and (2) a limited population is susceptible to colon carcinoma. The results are shown in Fig. 2.

For a limited population, the 13% line represents the best fit Weibull distribution to the data when three parameters are allowed to vary (A, k and λ; τ was fixed to be zero.) The χ^2 was minimized when $A = 13.5\%$, $k = 6.5$, and $\lambda = 90.5$. This provided an excellent fit to the data, as characterized by $\chi^2 = 99$ with 52 degrees of freedom giving a P-value=0.38.

The data is inconsistent with the hypothesis that 100% of the population is susceptible to developing colon carcinoma. The 100% line represents the best fit when only two parameters are allowed to vary (k and λ; while A was fixed to be 100%). This fit was not as good as characterized by $\chi^2 = 204$ with 53 degrees of freedom giving a P-value=10^{-10}. A visual inspection of the fit to the data, as seen in Fig. 2 is less convincing.

Finally, we tested whether weibull is consistent with the prostate carcinoma age-specific incidence data. We found that this captured the general shape of the prostate carcinoma age-specific

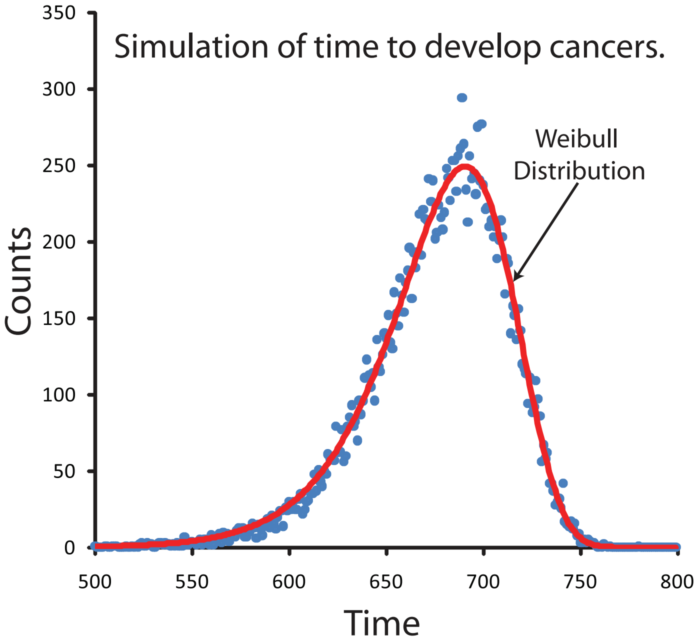

FIG. 1. We simulated the process of developing a tumor with a computer. Many (1000) random numbers were drawn from a parent distribution (a normal distribution with mean value 1000 and standard deviation of 100). This parent distribution represented the time for the many potential tumor progenitor cells in a tissue to develop into a tumor. The minimum of these many (1000) numbers was recorded, representing the first development of a tumor in the tissue and the age at which the subject was diagnosed with cancer. This process was repeated 20,000 times, representing a total population of 20,000 people. The data shown here represent a histogram of the age at which the 20000 people were first diagnosed with cancer. The general shape of the simulated data is not dependent upon the shape or parameters of the parent distribution, nor the total number in the population.

incidence curve, but the model did not completely encompass the data. The best fit model had parameters ($\tau = 34, A = 22200, k = 4.4$ and $\lambda = 40.3$). This gave a χ^2 value of 234 with 40 degrees of freedom. The probability of achieving such a high χ^2 value due merely to chance is 10^{-29}. Figure 3 shows a comparison between the data and model.

IV. DISCUSSION

The age-specific incidence data depends on several factors. The primary factor is the time for the cancer to develop, but a secondary factor is determined by screening and detection of the cancer. For instance, in the US it is recommended that people get a colonoscopy at age 50 to screen for colon cancer.[23,24] This results in a significant increase in colon cancer detected at age 50 and 51. Universal health care in the form of Medicare is available starting at age 65. This also leads to an increase in incidence observed at that age.

Several anomalies appear in the prostate carcinoma age-specific incidence data, as shown in Fig. 3. A significant increase occurs between the ages of 53 and 54, again between 61 and 62, and a third time between 64 and 65. This third increase also appears in the colon carcinoma data and is probably due to the beginning of eligibility for Medicare at age 65.

The colon carcinoma model predicts declining incidence of colon cancer after age 94, as seen in Fig. 2. Furthermore, it predicts an increasingly larger difference between a model in which everyone develops cancer and one in which a limited subset of the population develops cancer. Therefore, testing the model with data from ages 85-100 would provide a more convincing case that colon carcinoma only occurs in a limited subpopulation.

Because extreme value theory only relies on a few general assumptions, it may be applicable to other human chronic diseases. For instance, ischemic strokes occur when a blood vessel supplying the

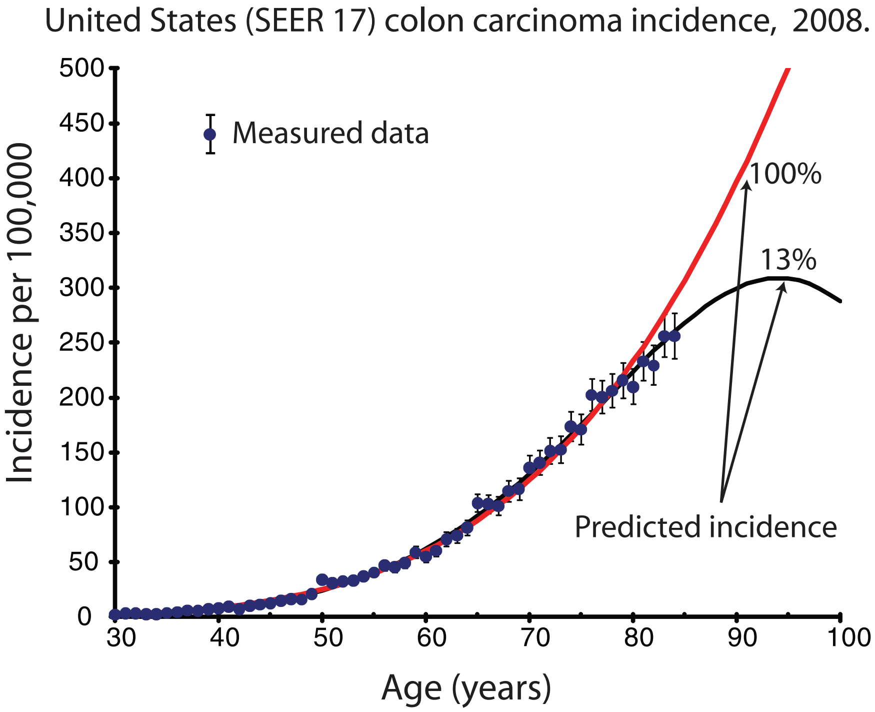

FIG. 2. This shows the age-specific incidence data for colon carcinoma in women during 2008 as recorded by the SEER-17 network of cancer registries. The error bars represent 95% confidence intervals. We used parameters for two different models: first, everyone (100%) in the population is susceptible to colon carcinoma, and second an unknown subset of the population is susceptible. We found the best fit when the subset is equal to about 13% of the population.

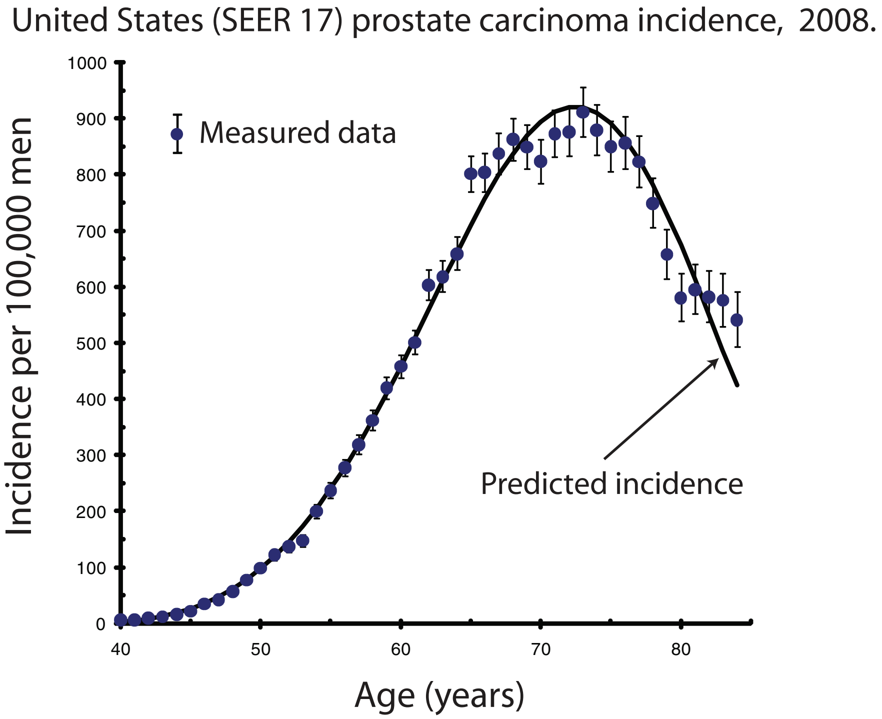

FIG. 3. The figure presents the age-specific incidence data for prostate carcinoma in men during 2008 as recorded by the SEER-17 network of cancer registries. The error bars represent 95% confidence intervals. The solid line depicts the best fit Weibull model.

brain become blocked. A plaque builds up on the artery wall through the process of atherosclerosis. A rupture in the plaque can lead to complete blockage of the artery. Atherosclerosis can occur at many sites, but one has an ischemic stroke only when the first rupture occurs. Thus, this process might be described by extreme value statistics.

V. CONCLUSION

We found that extreme value theory is consistent with the age-specific incidence of prostate and colon carcinomas. Furthermore, this implies that a limited subset of the population is susceptible to developing both diseases.

[1] D. Hanahan and R. A. Weinberg, Cell **100**, 57 (2000).
[2] D. Hanahan and R. A. Weinberg, Cell **144**, 646 (2011).
[3] B. Vogelstein and K. W. Kinzler, Trends Genet **9**, 138 (1993).
[4] E. R. Fearon, S. R. Hamilton, and B. Vogelstein, Science **238**, 193 (1987).
[5] T. Reya, S. J. Morrison, M. F. Clarke, and I. L. Weissman, Nature **414**, 105 (2001).
[6] B. A. Kohler, E. Ward, B. J. McCarthy, M. J. Schymura, L. A. G. Ries, C. Eheman, A. Jemal, R. N. Anderson, U. A. Ajani, and B. K. Edwards, J Natl Cancer Inst **103**, 714 (2011).
[7] P. Armitage and R. Doll, Br J Cancer **8**, 1 (1954).
[8] P. Armitage and R. Doll, Br J Cancer **91**, 1983 (2004).
[9] E. Farber, Cancer Res **44**, 4217 (1984).
[10] P. Armitage and R. Doll, Br J Cancer **11**, 161 (1957).
[11] S. H. Moolgavkar, Int J Epidemiol **33**, 1182 (2004).
[12] R. Meza, J. Jeon, S. H. Moolgavkar, and E. G. Luebeck, Proc Natl Acad Sci U S A **105**, 16284 (2008).
[13] T. Mdzinarishvili and S. Sherman, Cancer Inform **9**, 179 (2010).
[14] S. A. Frank, *Dynamics of Cancer: Incidence, Inheritance, and Evolution* (Princeton University Press, 2007) .
[15] S. A. Frank, Curr Biol **14**, 242 (2004).
[16] C. Harding, F. Pompei, E. E. Lee, and R. Wilson, Cancer Res **68**, 4465 (2008).
[17] G. Ritter, R. Wilson, F. Pompei, and D. Burmistrov, Toxicol Ind Health **19**, 125 (2003).
[18] F. Pompei and R. Wilson, Toxicol Ind Health **18**, 365 (2002).
[19] X. Zhang and R. Simon, Breast Cancer Res Treat **91**, 121 (2005).
[20] R. A. Fisher and L. H. C. Tippett, Proc. Cambridge Phil. Soc. **24**, 180 (1928).
[21] Surveillance, Epidemiology, and End Results (SEER) Program, “Seer*stat database: Incidence - seer 17,” (2010).
[22] L. Lasdon, A. Waren, A. Jain, and M. Ratner, ACM Transactions on Mathematical Software **4**, 34 (1978).
[23] U.S. Preventive Services Task Force, Ann Intern Med **137**, 129 (2002).
[24] U.S. Preventive Services Task Force, Ann Intern Med **149**, 627 (2008).