საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 19(193), №3, 2025
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 19(193), no. 3, 2025

Mathematics

Mathematical Modeling of Age-Specific Cancer Incidence Rates

Tengiz Mdzinarishvili[*], Revaz Shanidze[*], Mariam Osepashvili[*]

[*] *Kutaisi International University, Georgia*

(Presented by Academy Member Elizbar Nadaraya)

Abstract. Mathematical modelling of carcinogenesis attempts to explain the origins of cancer development at the individual and population levels. For interpretation age-specific cancer incidence rates across races of both sexes, the data of the Surveillance, Epidemiology, and End Results (SEER) Program are used. It is found that the beta model (BM), as well as the generalized beta model (GBM) fit very well the SEER age-specific incidence rates for 2017-2021, considering all cancer sites combined. The SEER data indicate that at the population level the hazard of getting a cancer for the female population is lower than for the male population. This statement is valid for all races as well as for the black and white populations. The weak point of these models is that they treat the entire population as a single homogeneous risk class. In this paper, we use a simple hypothesis of the dichotomous susceptibility to cancer in the population and give the interpretation of the observed data. The proposed approach is applied to modeling the cancer occurrence using the data collected in the SEER for all sites combined databases. We applied the conventional survival analysis formalism and obtained the equation connecting population hazard function (unconditional hazard function) with individual hazard function (conditional hazard function). The application of the regression analyses to the GBM shows that parameters of this model are statistically different, while the estimates of the cancer individual hazard functions appear visually are nearly the same. The obtained results may be considered as an argument for the carcinogenesis of multistage theory. © *2025 Bull. Georg. Natl. Acad. Sci.*

Keywords: mathematical modeling, statistics, cancer

Introduction

A mathematical modeling of the age distribution of cancer incidence rates results in a simple analytical function of the age t , $I(t)$ that can approximate observed values of cancer incidence rates and provides parameters of this function. The obtained model parameters can be further used to build rigorous statistical and/or biological models of cancer development (Mdzinarishvili, et al., 2009; Hiller et al., 2017).

Cancer is a disease of old age. However, contrary to this statement, it seems that cancer incidence levels off or even begins to fall after the age 75-80 (Arbeev, et al., 2005; Cox & Huber, 2007). This pattern appears almost universally across all cancer types.

Historically, three main hypotheses have emerged to explain this turnaround: 1) under-reporting of cancer incidence, 2) cellular mortality and 3) risk heterogeneity and mixing. The exact picture of how cellular senescence and apoptosis interact with carcinogenesis is complicated (Campisi, 2013; Rodier & Campisi, 2011; Cerella et al., 2016). It is thought that these two processes play crucial roles in the body's ability to prevent cancer. In fact, it is believed that escaping these two evolutionary fail-safes is a key step in the progression from cancer cell to tumor (Ershler & Longo, 1997; Su et al., 2015). The first epidemiological model to try to explicitly capture the impact of cellular mortality on carcinogenesis was proposed in (Pompei, & Wilson, 2001; Harding et al., 2008; Harding et al., 2012). According to this so called Beta Model (BM) for a given cancer incidence rate, we have the equation

$$I(t) = at^{k-1}(1-\beta t)\,. \qquad (1)$$

In this BM, at^{k-1}represents the incidence function of the k – stage multistage model (so a is a combined rate constant for limiting stage transitions) and βt is the probability that a cell and its linear descendants have become extinct from cellular senescence by time t, k is the number of limiting (slow) stages to produce the cancer, and β is an empirical term.

The modification of BM was proposed in (Mdzinarishvili, et. al., 2009). The authors used the generalized beta model (GBM) of carcinogenesis to fit data of the pancreatic and kidney cancers:

$$Ir(T) = c(bT)^{k-1}(1-bT)^{m-1}, \qquad (2)$$

where $T = (t - A)$, t is the age at cancer diagnosis; the incidence rate $Ir(T) = I(t)$; $b = 1/(B - A)$; A and B are the lower and upper age limits of cancer development in years, respectively (we assume $A = 0$ and $B = 100$)*;* c is a generalized rate constant; k is the number of stages required to initiate cancer; m is the degree of decrease in cancer incidence rates due to the probability of stem cell death (for whatever reason). The equation (2) can be rewritten as

$$I(t) = c\left(\frac{t-A}{B-A}\right)^{k-1}\left(\frac{B-t}{B-A}\right)^{m-1}. \qquad (3)$$

A feature of both these models (BM and GBM) is that the predicted time of peak incidence of these models is independent of the transition rate parameter a and generalized rate constant c, correspondingly. Thus, both these models predict that an increase of cancer incidence at a younger age would lead to overall higher rate in the population and turn around at the older age.

The weak point of these models is that they treat the entire population as a single homogeneous risk class. This is problematic considering the great amount of evidence that genetic variations can be predictors for increased risk (Mdzinarishvili & Sherman, 2014; Gsteiger & Morgenthaler, 2008). This idea of genetic variation being a cause of risk heterogeneity leads to the third explanation for the turnaround at old age (for the details see Hiller et al., 2017; Mdzinarishvili & Sherman, 2014).

In this paper, we use a hypothesis of the dichotomous susceptibility to cancer in the population and give the interpretation of the observed data. The proposed approach is applied to modeling cancer occurrence using data, collected in the Surveillance, Epidemiology, and End Results (SEER) on the all-sites-combined databases. The obtained new results may help in understanding of the carcinogenesis process.

Materials and Methods

The SEER databases provide delay factors specific to a cancer site, registry, age group, race, and year of diagnosis. In our analysis for age-specific incidence rates, we use all cancer sites combined SEER delay-adjusted Incidence rates by age at diagnosis, 2017-2021, with their standard errors. We do not take into account the birth cohort and time period effects (Mdzinarishvili et al., 2009), assuming that all these effects are negligible. Bellow, we use the data from SEER*Explorer, 2025.

In this work, we use the concepts and designations of the survival analysis that were adapted in (Mdzinarishvili & Sherman, 2013; Mdzinarishvili & Sherman, 2014) for purposes of carcinogenic modeling. By $S(t)$ we denote a conditional survival function that an individual "survives" from getting a particular type of cancer at the age t, giventhat this individual belongs to the pool of individuals susceptible to cancer, and we called $S(t)$ the individual survival function. Analogously, we denote $h(t)$and $f(t)$the conditional hazard function and the conditional probability density function, correspondingly. We also called $h(t)$ the individual hazard function (as well as the individual cancer presentation function) and $f(t)$ – the probability density function of individual cancer presentation. According to the conventional survival analysis formalism, $S(t)$, $h(t)$ and $f(t)$ are related in the following way (Mdzinarishvili & Sherman, 2013):

$$h(t) = \frac{f(t)}{S(t)}, \quad (4)$$

$$f(t) = -\frac{dS(t)}{dt}, \quad (5)$$

$$S(t) = exp\left(-\int_0^t h(z)dz\right) = exp[-H(t)], \quad (6)$$

$$H(t) = \int_0^t h(z)dz, \quad (7)$$

where $H(t)$ is the cumulative individual hazard function.

We denote by $S_U(t)$ an unconditional survival (or cancer resistance) function showing that an individual, randomly chosen from the population, did not experience the event (cancer presentation) at the age t (i.e. this individual did not develop cancer up to that age). We called $S_U(t)$ the population (unconditional) cancer resistance function. Analogously, we denoted the unconditional cancer hazard function (or population cancer hazard function) by $h_U(t)$.

According to (Mdzinarishvili & Sherman, 2013), $S_U(t)$ and $S(t)$ are related as follows:

$$S_U(t) = (1-p)\cdot 1 + pS(t) = 1 - p + pS(t) \quad (8)$$

and

$$S(t) = \frac{1}{p}[S_U(t) + p - 1], \quad (9)$$

where p is the probability that a randomly chosen individual is susceptible to cancer and $1-p$ is the probability that this individual is resistant to cancer. Note, p can also be considered as the relative size of the pool of the individuals susceptible to cancer.

The unconditional cancer hazard function, $h_U(t)$, showing that an individual, randomly chosen from the whole population, gets cancer at the age t can be presented in the following way:

$$h_U(t) = \frac{ph(t)}{p+(1-p)\exp[H(t)]}. \quad (10)$$

The conditional cancer hazard function, $h(t)$, which shows that an individual, randomly chosen from the pool of individuals susceptible to cancer, is diagnosed with cancer at the age t is presented as:

$$h(t) = \frac{h_U(t)}{H_{UO}(t) - H_U(t)}, \quad (11)$$

where:

$$H_U(t) = \int_0^t h_U(z)dz \quad (12)$$

is the cumulative unconditional cancer hazard function, and

$$H_{UO}(t) = \int_0^\infty h_U(z)dz \quad (13)$$

is the entire cumulative unconditional cancer hazard.

In cancer registries, the cancer incidences and the size of population are usually tabulated with five-year age intervals. SEER presents estimates of the age-specific incidence rates (unconditional hazard function $h_U(t_i)$), where the age $t_i (i = 1,2,\dots,n)$ is a discrete variable presenting the corresponding n successive age intervals mid points. We called the conditional cancer hazard, $h(t_i)$, as the individual cancer hazard and unconditional cancer hazard, $h_U(t_i)$, as the population level hazard.

Results

From (11) it follows that an empirical estimate (denoted by sign'^') of the conditional cancer hazard function, $\hat{h}(t_i)$ can be obtained by the following formula

$$\hat{h}(t_i) = \frac{\widehat{h_U}(t_i)}{\widehat{H_{UO}}(t_i) - \widehat{H_U}(t_i)} \quad (i = 1,2,\dots,n), \quad (14)$$

where estimates $\widehat{h_U}(t_i)$, $\widehat{H_{UO}}(t_i)$ $\widehat{H_U}(t_i)$ and their standard errors are obtained from SEER data by standard procedures. For all races we give the observed SEER data.

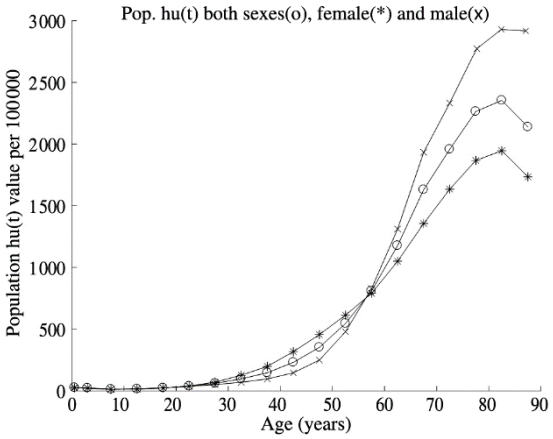

Fig. 1. Estimates of the population hazard functions.

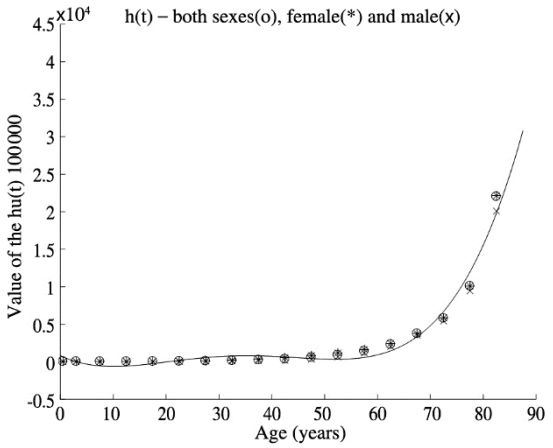

Fig. 2. Estimates of the individual hazard functions.

Figure 1 presents the estimates of the cancer population hazard functions, $\widehat{h_U}(t_i)$, for all cancer sites combined and all races, based on the SEER age-specific incidence rates from 2017 to 2021. The small circles (o) denote SEER incidence rates for both sexes at the midpoints of the first 19 time intervals presented in the first column of the Table (SEER*Explorer:). Standard errors are too small and are not presented in Fig. 1. The asterisk (*) denotes SEER incidence rates for female population and (x) for male population. Regression analyses for BM and GBM shows that parameters of these models are statistically different, while in Fig. 2 the estimates of the cancer individual hazard functions $\hat{h}(t_i)$, are visually very close. The black line indicates 4th order polynomial approximation.

Calculations show (results are not presented) that the observed age-specific incidence data have good fitting for GBM lines. We calculated the coefficient of determination R^2 for considered populations and found that this statistics is very close to 1. This means that the best fit curves match the data with nearly no scatter for considered populations. It should be noted that $R^2 \approx 1$ indicates that the curve came very close to the data points, but does not imply that the fit is sensible in other ways (Motulsky & Christopoulos, 2004).

Conclusion

A simple hypothesis of the dichotomous susceptibility to cancer in the population is used for interpretation of the observed data. The proposed approach is applied for modeling the cancer occurrence, using data collected in the SEER all-sites-combined databases. Using the conventional survival analysis formalism gives the equation connecting population hazard function (unconditional hazard function) with individual hazard function (conditional hazard function). Application of the regression analyses to the GBM shows that parameters of this model are statistically different, while the estimates of the cancer individual hazard functions are very close. The obtained results may be considered as an argument of the carcinogenesis multistage theory.

Acknowledgements

The work was carried out at the Kutaisi International University and supported by the Shota Rustaveli National Science Foundation of Georgia: STEM grant 22-1714.

მათემატიკა

კიბოთი ავადობის მათემატიკური მოდელირება ასაკის მიხედვით

თ. მძინარიშვილი*, რ. შანიძე*, მ. ოსეფაშვილი*

* *ქუთაისის საერთაშორისო უნივერსიტეტი, საქართველო*

(წარმოდგენილია აკადემიის წევრის ე. ნადარაიას მიერ)

კანცეროგენეზის მათემატიკური მოდელირება მიზანმიმართულია ახსნას კიბოს წარმოშობა და განვითარება ინდივიდუალურ და პოპულაციურ დონეზე. ყველა რასის, ორივე სქესის, ქალისა და მამაკაცის კიბოს ასაკობრივი სპეციფიკური შემთხვევების ინტერპრეტაციისთვის გამოიყენება SEER (The Surveillance, Epidemiology, and End Results) პროგრამის დაკვირვებათა მონაცემები. აღმოჩნდა, რომ ე.წ. ბეტა მოდელი (ბმ), ისევე როგორც განზოგადებული ბეტა მოდელი (გბმ), ძალიან კარგად ესადაგება SEER-ის ასაკობრივად სპეციფიკურ შემთხვევების მაჩვენებლებს 2017-2021 წლებში, კიბოს ყველა სახეობის კომბინირებულ მონაცემებს. SEER მონაცემებით მიღებულია, რომ პოპულაციის დონეზე, კიბოთი დაავადების რისკი ქალებში უფრო დაბალია, ვიდრე მამაკაცებში. ეს მტკიცება მართებულია ყველა რასისთვის, ასევე შავკანიანი და თეთრკანიანი მოსახლეობისთვის. წინამდებარე ნაშრომში გამოყენებულია პოპულაციაში კიბოს დაავადების მიმართ დიქოტომიური მგრძნობელობის მარტივი ჰიპოთეზა და შემოთავაზებულია დაკვირვებული მონაცემების ინტერპრეტაცია. შემოთავაზებული მიდგომა გამოიყენება კიბოს შემთხვევების მოდელირებისთვის SEER-ის გაერთიანებულ მონაცემთა ბაზებში შეგროვებული მონაცემების მიხედვით. ჩვენ გამოვიყენეთ გადარჩენის ტრადიციული ანალიზის ფორმალიზმი და მივიღეთ განტოლება, რომელიც აკავშირებს პოპულაციის საფრთხის ფუნქციას (უპირობო საფრთხის ფუნქციას) ინდივიდუალურ საფრთხის ფუნქციასთან (პირობით საფრთხის ფუნქციასთან). რეგრესიული ანალიზის გბმ-ზე გამოყენება აჩვენებს, რომ ამ მოდელის პარამეტრები სტატისტიკურად განსხვავებულია, მაშინ როდესაც კიბოს ინდივიდუალური საფრთხის ფუნქციების შეფასებები ვიზუალურად თითქმის ერთნაირია. მიღებული შედეგები შეიძლება ჩაითვალოს კანცეროგენეზის მრავალსაფეხურიანი თეორიის არგუმენტად.

REFERENCES

Arbeev, K., Ukraintseva, S., Arbeeva, L., & Yashin A. (2005). Decline in human cancer incidence rates at old ages: Age-period-cohort considerations. *Demographic Research*, *12*(11), 273-300. doi: 10.4054/DemRes.2005.12.11.

Campisi, J. (2013). Aging, cellular senescence, and cancer. *Annu Rev Physiol.,* 75, 685-705. doi: 10.1146/annurev-physiol-030212-183653.

Cerella, C., Grandjenette, C., Dicato, M., & Diederich, M. (2016). Roles of apoptosis and cellular senescence in cancer and aging. *Curr Drug Targets*, *17*(4),405-415. Doi.org / 10.2174/1389450116666150202155915

Cox, Jr L., & Huber, W. (2007). Symmetry, identifiability, and prediction uncertainties in multistage clonal expansion (MSCE) models of carcinogenesis. *Risk Anal.*, *27*(6), 1441-1453. doi: 10.1111/j.1539-6924.2007.00980.x.

Ershler, W., & Longo D. (1997). Aging and cancer: Issues of basic and clinical science. *JNCI J Natl Cancer Inst, 89*(20), 1489-1497. doi: 10.1093/jnci/89.20.1489.

Harding, C., Pompei, F., Lee, E., & Wilson, R. (2008). Cancer suppression at old age. *Cancer Res., 68(*11), 4465-4478. doi: 10.1158/0008-5472.CAN-07-1670.

Harding, C., Pompei, F., Wilson, R. (2012, March 1). Peak and decline in cancer incidence, mortality, and prevalence at old ages. *Cancer.* doi: 10.1002/cncr.26376. Epub 2011 Sep 22.

Hiller, J., Vallejo, C., Betthauser, L. and Keesling, J. (2017). Characteristic patterns of cancer incidence, *Prog Biophys Mol Biol.* March 1. 24, 41-48.

Mdzinarishvili, T., & Sherman, S. (2013). Basic equations and computing procedures for frailty modeling of carcinogenesis: Application to pancreatic cancer data. *Cancer Informatics, 12*(67). doi: 10.4137/CIN.S8063.

Mdzinarishvili, T., & Sherman, S. (2014). Heuristic modeling of carcinogenesis for the population with dichotomous susceptibility to cancer: A pancreatic cancer example. *PLoS ONE, 9*(6). doi.org/10.1371/journal.pone.0100087.

Mdzinarishvili, T., Gleason, M., Kinarsky, L., & Sherman, S. (2009). A generalized beta model for the age distribution of cancers: application to pancreatic and kidney cancer. *Cancer Inform., 7,* 183-197. doi. org/10.4137/CIN.S30.

Motulsky, H., & Christopoulos, A. (2004). Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting. New York.

online edn, Oxford Academic, 31 Oct. 2023). Doi.org / 10.1093/oso/ 9780195171792.001.0001, accessed 9 June 2025.

Pompei, F., & Wilson, R. (2001). Age distribution of cancer: The incidence turnover at old age. *Human and Ecological Risk Assessment, 7*(6), 1619-1650.

Rodier, F., & Campisi, J. (2011) Four faces of cellular senescence. *JCB, 192*(4), 547-556, 2. doi: 10.1083/jcb.201009094. Epub 2011 Feb 14.

SEER*Explorer: An interactive website for SEER cancer statistics [Internet]. Surveillance Research Program, National Cancer Institute; 2025 Apr 16. [cited 2025 Jun 18]. Available from: https://seer.cancer.gov/statistics-network/explorer/

Su, Z., Yang, Z., Xu, Y., Chen, Y., & Yu., Q. (2015). Apoptosis, autophagy, necroptosis, and cancer metastasis. *Mol. Cancer, 14*(48), 1-14. doi: 10.1186/s12943-015-0321-5.

Received May, 2025