To appear in Serdica Math. J. **44** (2018)

Serdica
Mathematical Journal

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

SOME MATHEMATICAL PROBLEMS IN CANCER RESEARCH

Mariia Beliaeva

Communicated by M. Savov

Abstract. Some mathematical models used in cancer research are considered. Some mathematical mistakes made in those models are analyzed. Also a new version of the well-known multistage model of carcinogenesis is presented.

1. Introduction. It often happens that serious mathematicians look down on some applied problems believing that mathematics used there is primitive and is not worth their attention. On the other hand, specialists in that non-mathematical domain consider mathematical models as a decoration for their theories and do not care much about mathematical rigor and correctness. This phenomenon is a sort of *Mechanitis* — "the occupational disease of one who... believes that a mathematical problem, which he can neither solve nor even formulate, can readily be answered, once he has access to a sufficiently expensive

2010 *Mathematics Subject Classification:* 92C50, 97M60, 62P10.
Key words: cancer mathematical models, mistakes, multistage model, carcinogenesis, Bad Luck theory.

machine" [20], with the only difference being that instead of machines a couple of formulas are believed to be enough.

This situation provides a rich pasture for a mathematician who is curious enough to stick his head into a new domain, and stubborn enough to learn strange terms and ideas which are used there. First, one can have fun observing remarkable mistakes in such quantities that one could never meet in mathematical environment. Let us look at some examples:

a. "... a constant "extrinsic mortality" rate of 0.1 per year implies that overall lifespan cannot be extended beyond $1/0.1 = 10$ years ..." [33];

b. "... the probability of a particular cell mutating is x ... the probability of having a colony of n mutated cells somewhere in a tissue containing N cells is Nx^n" [15]. The authors without batting an eye get this probability equal to 3.2;

c. The parameters chosen for Weibull distribution imply $\int p(t) = 20223$ where $p(t)$ is the probability density function [27];

d. "Even when stem-cell mutations occur at random, the initiation ... of a cancer cannot be viewed as a random process" [12].

Second, developing a more or less reasonable model in a new domain is a challenge, which requires not only mathematical knowledge but also a large volume of common sense and some impudence to start playing on a foreign turf.

It seems that the very same thing happens now in cancer mathematics. Inventing various models of cancer initiation and progression is very fashionable today. But for some reason many respected cancer specialists develop mathematical models without professional mathematical assistance. On the basis of their amateurish creations they sometimes make serious conclusions, which are supposed to give answers to fundamental problems of the nature of cancer and thus determine general directions of further research.

2. The three problems. We can name at least three fundamental problems in cancer research, such that attempts to solve each one of them led to creating cancer mathematical models (CMM). Historically, the first of those models was developed in order to explain regularities discovered in incidence data [4]. In the 1950s Armitage and Doll [1] noticed that the logarithm of the death rate increased proportionally to the logarithm of age, but about six times as rapidly; in other words, the death rate increased proportionally to the sixth power of age. They concluded that a cancer cell was the end-result of seven successive mutations. This multistage theory (MT) of carcinogenesis has undergone several changes afterwards and gave rise to a variety of different theories owing to new data and new ideas emerging. The lifespan has grew up, statistics improved, and

the data for age groups above 75 years became available. It turned out that the incidence rate that replaced the death rate in MT did not fit that log-log curve [27] as its increase was slowing down in old ages.

Some researchers attempted to explain those incidence curves as accumulation of only three mutations [14]. Other propositions include extreme value theories [17],[27], population genetics model [24] and many other evolutionary models [5], and even such exotic ones as kinetic+game models [6] or Landau model of the second order phase transition [28]. The problem is still being debated.

The second problem for which CMM are widely used is to understand the role of intrinsic vs extrinsic factors in cancer initiation. The sensational Bad Luck Theory report [29] by two prominent scientists became the major event in this field two years ago. The authors argue that intrinsic stochastic effects play the main role in tumor initiation, hence primary prevention (vaccination, altering lifestyle, environmental control) is unlikely to prevent a large subset of cancers, especially hereditary ones. On the contrary, secondary prevention (early detection and treatment), in their view, has to be the major focus. This result was severely criticized from biological point of view by many cancer specialists (for bibliography see [32]) but the disputants waste their time: the underlying mathematical model is incorrect [7]. Its main idea is that as coefficient of correlation R between two variables, namely number of stem cell divisions and cancer incidence, is equal to 0.804, hence $R^2 \approx 2/3$ of actual cancer incidence can be explained by stem cells random divisions, i.e. by bad luck.

Actually, the authors calculate correlation between logarithms of two variables and interpret it as correlation between the variables themselves. This is a gross mathematical mistake as correlation between non-linear functions is not equal to correlation between the arguments. The real coefficient of correlation between stem cells divisions number and cancer incidence is 0.97, so according to the Bad Luck logic the percentage of unpreventable cancers should be more than 0.9. But it is well known that correlation does not imply causation, hence it does not explain anything and the conclusion is unfounded. The mathematical level of [29] can be illustrated also by "extra risk score", which the authors develop as an index of "how high risk is relative to division number", but calculate it not as ratio but as a product of these two values. The authors have clarified this as "It may seem intuitive to multiply rather than add logarithms" [31]. And using machine learning to split a single set of thirty values of this index into two subsets of negative and positive numbers, which can be successfully accomplished by means of ones eyes, is obviously the example of decoration that we have discussed

above.

Meanwhile the recent research [19] suggests that each of risk factors such as alcohol, smoking, Body Mass Index, lack of physical activities, diet, etc increases colorectal cancer risk by 35%, which means that at least this particular cancer is preventable for many people. This result does not contradict the high correlation calculated in [29] but is obviously incompatible with the famous Bad Luck conclusion.

The third problem is about the hypothesis of frailty which means that "a fraction of population is either exclusively at risk, or at vastly increased risk compared with the general population" [22]. Indeed, why do not the majority of people get cancer? Is it true that everybody gets cancer if he lives long enough, or some people are immune to it? An elegant idea was proposed in [27]. The authors suggest that age-specific incidence follows the extreme value (Weibull) distribution if cancer is diagnosed as soon as the first of many potential tumor cells develops into a tumor. Choosing the distribution parameters so that the cumulative distribution function fits the incidence curve they get that only 13.5% of population is susceptible to colon and only 22% — to prostate carcinomas. Unfortunately the authors use wrong formulas for Weibull distribution as well as for statistical criteria [8] and thus this conclusion is unfounded too.

Nevertheless, the main problem of that research is not the incorrect equalities but the hidden assumptions which, as it usually happens, are neither discussed nor even mentioned. In fact, to use Weibull distribution here one has to assume a) that all potential tumor cells in the population behave equally, i.e. that cancer start times in all bodies (and in all cells in any specific body) have the same distribution, and b) that for different cells in any body those times are independent. Both are questionable. Although the role of immunity in cancer development is not fully understood there is a lot of evidence suggesting that immune system does prevent cancer [9],[10]. So cancer incidence in a particular person depends on the state of his/her immune system. Obviously these states are different in different people because of different heredity or/and of different lifestyles, environmental and occupational exposures, current and past infections, etc. Thus the first assumption is not valid. As for the second one, there is some evidence that stem cells interact with each other by means of transforming the environment (characterized by extracellular matrix [16]) around themselves. If so, this interaction may be positive (the mutated cell transforms the environment in the way that facilitate mutations in the nearby cell) as well as negative (the mutated cell impedes the others mutation). The idea of dependence is based on the general fact that there is no real random process whose values are all inde-

pendent of each other. Time of developing cancer in a cell is a random process, thus start times for some cells in a tissue should be dependent.

All three problems are not only of academic interest but also of great practical importance. Knowing what factors influence cancer initiation the most (the second problem) as well as understanding whether all the population or only some part of it is susceptible to cancer (the third problem) would help to optimize cancer research funding, screening schemes, and cancer prevention, while solving the first problem would help with the other two. The cancer research community uses mathematical methods widely but unfortunately does not have a habit of rejecting the incorrect solutions, which is a necessary part of applying mathematics in any domain. The general public has a great interest to the progress in this field, and because of that some weakly founded or unfounded results become sensations shared widely by mass media, which in their turn influence the decision makers. So examining mathematical correctness of existing CMMs in order to avoid further mistakes is of key importance.

3. Multistage model and its correction. The MT model, which was discussed in the previous Section, "has been a pillar of the mathematical and statistical study for decades" [18]. Its main assumption is that cancer in a tissue starts when at least one of its cell lineages achieves malignant state after having undergone M sequential transformations (*driver mutations*). Time intervals between successive mutations are assumed to be independent random values following exponential distributions with densities $\lambda_i \exp(-\lambda_i x)$, $i = 1, \dots, M$. λ-s are usually called *transition rates* [23]. The cumulative probability $p(T)$ of getting cancer during the first T years of life is [2]

$$p(T) = p\{\xi_M < T\} = 1 - \prod_{j=1}^{M} \lambda_j \times \sum_{i=1}^{M} \frac{1}{\lambda_i C_i^{(M)}} e^{-\lambda_i T} \tag{1}$$

where $\xi_M = t_1 + t_2 + \cdots + t_M$, $C_i^{(M)} = \prod_{j=1, j\neq i}^{M} \Big(\lambda_j - \lambda_i\Big)$.

Cancer statistics though is not about cumulative probability but about **cancer incidence rate**, which is "the number of new cancers of a specific site/type occurring in a specified population during a year, usually expressed as the number of cancers per 100,000 population at risk

$$Incidence\ rate = (New\ cancers/Population) \times 100,000".$$

This is the definition from the *Surveillance, Epidemiologe, and End Results Program*, a US population-based registry that records all cancers regardless

of clinical treatment [36] and covers approximately 28% of the US population. The National Cancer Institute [34] treats cancer incidence as the normalized number of new cases of cancer too. Thus incidence rate is the 100000-fold sample estimate of conditional probability for a person to get cancer at age T given that he didn't get it before.

Not having enough data about incidence rate in those times, Armitage and Doll were studying mortality rate, which had the similar sense: it is "a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time" [37], thus it is also the sample estimate of conditional probability for a person to die at age T, multiplied by some constant (1000 according to [37]). Indeed, the already dead people are removed from the future statistics, which means that incidence rate takes into account only those who are still alive.

The conditional probability for one cell lineage to get M mutations during T years given that less than M mutations happened to it during the first (T-1) years is

$$p_c(T) = P\{m_T \geq M | m_{T-1} < M\} = \frac{P\{m_T \geq M, m_{T-1} < M\}}{P\{m_{T-1} < M\}}, \tag{2}$$

where m_T is the number of mutations the cell underwent during first T years of life.

Equality (2) is equivalent to

$$p_c(T) = \frac{P\{T-1 \leq \xi_M \leq T\}}{P\{\xi_M > T-1\}} = \frac{P\{\xi_M \leq T\} - P\{\xi_M \leq T-1\}}{P\{\xi_M > T-1\}}. \tag{3}$$

After a simple transformation we get

$$p_c(T) = 1 - \frac{1-p(T)}{1-p(T-1)}, \tag{4}$$

where $p(T)$, $p(T-1)$ are calculated as in (1). Similarly, the conditional probability for a tissue to get cancer at age T is

$$p_c^{(tiss)}(T) = 1 - \frac{1-P\{\tau \leq T\}}{1-P\{\tau \leq T-1\}}, \tag{5}$$

where τ is the time of cancer start in the tissue and $P\{\tau \leq T\}$ is the cumulative probability for the tissue to get cancer during the first T years of life. Assuming independence of mutation processes in different cells we get

$$P\{\tau \leq T\} = 1 - \Big(1 - p(T)\Big)^{N_{cells}}, \tag{6}$$

where N_{cells} is the number of cells in a tissue. Finally,

$$p_c^{(tiss)}(T) = 1 - \Big(\frac{1-p(T)}{1-p(T-1)}\Big)^{N_{cells}}. \tag{7}$$

Figure 1 shows the age-incidence rate curve (the SEER data for colon and rectum cancers) and its approximation (6) for $M = 6$. One can see that the exponential curve diverges from the incidence one after 75 years because of slowing of incidence increase that we discussed in Section 2.

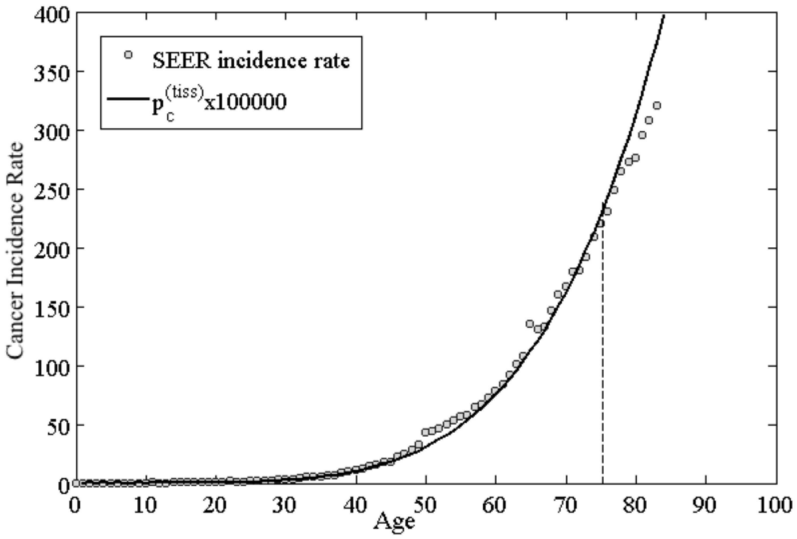

Fig. 1. Approximation of incidence rate by conditional probability (7)

Armitage and Doll noticed that the mortality rate of many cancers increased as sixth (fourth to sixth in [3]) power of age. Using that

$$p(T) \approx \frac{\lambda_1 \cdots \lambda_M}{M!} T^M \tag{8}$$

([21], quoted in [2]) they concluded that M may be equal to seven (around five to seven in [3]). This was the beginning of MT model. Nowadays some groups of researchers are dissatisfied with this model. Soto and Sonneschein claim that multistage theory "should be dropped and replaced" [26]. They propose instead their Tissue Organization Field Theory (TOFT) the main idea of which can be translated to the mathematical language as "all processes in the human body are interdependent and carcinogenesis is not an exception". Rozhok and De Gregory

[25] make a step in this direction linking transition rates with fitness, which is a decreasing function of age. The latter idea, while explaining the old age incidence decline, seems to be inconsistent with common sense: it implies that the younger and the healthier is the organism the more likely it is to get the disease. It seems more logical to treat the transition rates as functions of immunity state, which increases in the childhood and decreases in the old age, see Fig. 2.

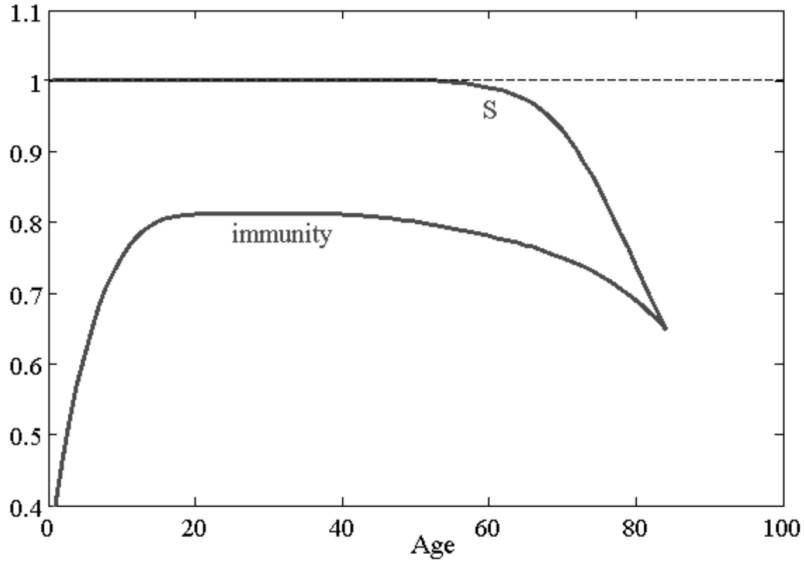

Fig. 2. Immunity and S curves used in nhP model

Indeed, there is a lot of evidence that immune system can recognize and destroy malignantly transformed cells [9, 11]. It is well known also that people with high-level immunity not only recover faster when they get common diseases, e.g. the cold, but also fall ill less often as compared to those with low immunity. With respect to cancer the lesser susceptibility means that either the number M of necessary mutations is less in people with low immunity or the transition rates are higher. We assume that transition rates are inversely proportional to immunity; rates for all M mutations at that are supposed to be equal for simplicity. In this case carcinogenesis becomes a non-homogeneous Poisson process with the transition rate $\lambda(t)$.

The second assumption that we make in our model is that life processes slow in the old age. Thus we multiply the transition rate $\lambda(t)$ by function $S(t)$ declining with age see Fig. 2. Now the cumulative probability for a cell lineage

to reach a malignant state during first T years of life is [35]

$$p_{nhP}(T) = \frac{\left[\lambda_{nhP}(T)\right]^M e^{-\lambda_{nhP}(T)}}{M!}, \tag{9}$$

where

$$\lambda_{nhP}(T) = \int_0^T \lambda^*(t)dt,$$

$\lambda^*(t) = \lambda(t) \times S(t)$. Finally, for tissue with N_{cells} cells we get

$$p_{nhP}^{(tiss)}(T) = 1 - \Big(\frac{1 - p_{nhP}(T)}{1 - p_{nhP}(T-1)}\Big)^{N_{cells}}. \tag{10}$$

Figure 3 shows that this model explains the incidence rate old age slowing.

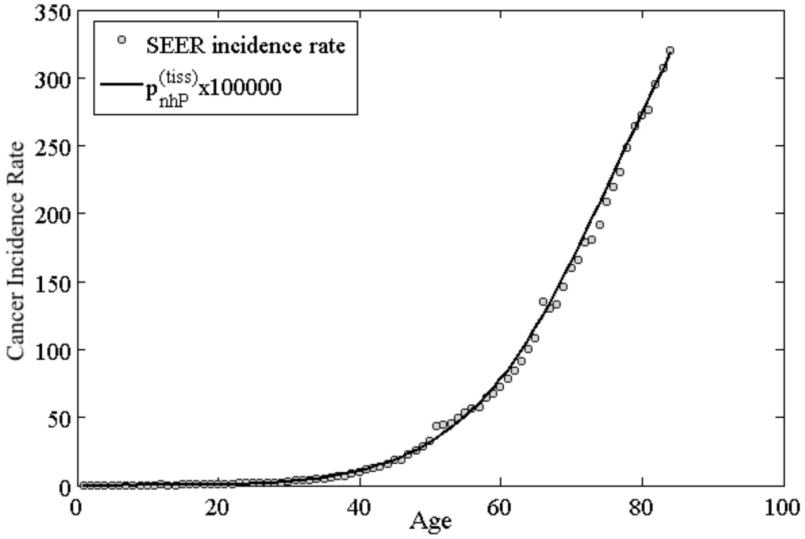

Fig. 3. Approximation of incidence rate by nhP probability (10)

4. Conclusions.

1. Mathematical community should pay closer attention to the attempts to use mathematics in cancer research.

2. A professional mathematician's assistance would help cancer specialists to avoid many mistakes and save them time and effort.

3. Decreasing of immunity together with slowing up of life processes with age can explain the cancer incidence curves old age behaviour.

The non-homogeneous model described in Section 3 is a first simple step towards the statistical estimation of how extrinsic factors affect the cancer incidence rate. Our next goal is to take into account the correlation between mutation processes in different cells.

Acknowledgements. I'm grateful to M. Mitrofanov for helpful discussion and to A. Rozhok for clarifying some aspects of mutation process for me.

REFERENCES

[1] P. Armitage, R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **8**, *1* (1954), 1–12.

[2] P. Armitage. A note on the time-homogeneous birth process. *J. Roy. Statist. Soc. B* **15** (1953), 90–91.

[3] P. Armitage. Multistage Models of Carcinogenesis. *Environmental Health Perspectives* **63** (1985), 195–201.

[4] C. S.-O. Attolini, F. Michor. Evolutionary Theory of Cancer. *Ann. N. Y. Acad. Sci.* **1168**, *1* (2009), 23–51.

[5] N. Beerenwinkel et al. Cancer Evolution: Mathematical Models and Computational Inference. *Syst Biol.* **64**, *1* (2015), 1-25.

[6] N. Bellomo, M. Delitala. From the mathematical kinetic, and stochastic game theory to modelling mutations, onset, progression and immune competition of cancer cells. *Physics of Life Reviews* **5** (2008), 183–206.

[7] M. Beljaeva. Bad Luck theory from the mathematicians point of view. `https://www.researchgate.net/publication/296696978`, 2016, 1–7.

[8] M. Beljaeva. Weibull Distribution Theory of cancer age-specific incidence from the mathematician's point of view. `https://www.researchgate.net/publication/310326046`, 2016, 1–6.

[9] A. Corthay. Does the Immune System Naturally Protect Against Cancer? *Frontiers in Immunology*, **5** (2014), 197, 8 pp.

[10] M. V. Dhodapkar. Personalized Immune-Interception of Cancer and the Battle of Two Adaptive Systems – When Is the Time Right? *Cancer Prev. Res. (Phila)* **6**, *3* (2013), 173–176.

[11] G. P. Dunn et al. Cancer immunoediting: from immunosurveillance to tumor escape. *Nature Immunology* **3** (2002), 991–998.

[12] M. Kelly-Irving, C. Delpierre, P. Vineis. Beyond bad luck: induced mutations and hallmarks of cancer. *The Lancet Oncology* **18**, *8* (2017), 999–1000.

[13] W. Feller. An Introduction to Probability Theory and Its Applications, vol. 2., 3nd. Ed. New Jersey, US, J. Wiley, 1968.

[14] J. C. Fisher. Multiple-mutation theory of carcinogenesis. *Nature* **181** (1958), 651–652.

[15] J. C. Fisher, J. H. Hollomon.A hypothesis for the origin of cancer foci. *Cancer* **4** (1951), 916–918.

[16] F. Gattazzo et al., Extracellular matrix: A dynamic microenvironment for stem cell niche. *Biochim Biophys Act* **1840**, *8* (2014), 2506–2519.

[17] G. R. Haynatzky et al. A new statistical model of tumor latency time. *Mathematical and Computer Modelling* **32** (2000), 251–256.

[18] J. Hiller, J. Keesling. Asymptotic Relative Risk Results from a Simplified Armitage and Doll Model of Carcinogenesis. *Bulletin of Mathematical Biology* **80**, *3* (2018), 670–686.

[19] G. Ibanez-Sanz et al. Risk Model for Colorectal Cancer in Spanish Population Using Environmental and Genetic Factors: Results from MCC-Spain study. *Sci. Rep.* **7**, 43263 (2017), 1–11.

[20] B. Koopman, C. Hitch. Fallacies in Operations Research. Operations Research, **4**, *4* (1956), 422–430.

[21] O. Lundberg. On Random Processes and their Application to Sickness and Accident Statistics. Uppsala, Almqvist & Wicksells, 1940.

[22] S. H. Moolgavkar. Commentary: Frailty and heterogeneity in epidemiological studies. *International Journal of Epidemiology* **44**, *4* (2015), 1425–1426.

[23] S. H. Moolgavkar. Theory of Carcinogenesis and the Age Distribution of Cancer in Man. *J. Natl Cancer Inst.* **61**, *1* (1978), 49–52.

[24] M. A. Nowak et al. Evolutionary dynamics of tumor suppressor gene inactivation. *Proc. Natl. Acad. Sci. USA* **101** (2004), 10635–10638.

[25] A. I. Rozhok, J. De Gregory. Towards an evolutionary model of cancer: Considering the mechanisms that govern the fate of somatic mutations. *PNAS* **112**, *29* (2015).

[26] C. Sonnenschein, A. M. Soto. Somatic mutation theory of carcinogenesis: why it should be dropped and replaced. *Mol. Carcinog.* **29**, *4* (2000), 205–211.

[27] L. Soto-Ortiz, J. P. Brody. (2012) A theory of the cancer age-specific incidence data based on extreme value distributions. *AIP Advances* **2**, 011205, 6 pp.

[28] V. G. Soukholovsky et al. The population dynamics of cancer incidence: the model of a second-order phase transition. *Biophysics* **60**, *4* (2015), 639–646.

[29] C. Tomasetti, B. Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, *6217* (2015), 78–81.

[30] C. Tomasetti, B. Vogelstein. Musings on the theory that variation in cancer risk among tissues can be explained by the number of divisions of normal stem cells. `arXiv:1501.05035`, 2015, 17 pp.

[31] C. Tomasetti, B. Vogelstein. Response. *Science* **347**, *6223* (2015), 729–731.

[32] T. C. Wang, E. Szabo. Implications of the "bad luck" explanation of cancer risk for the field of cancer prevention. *Cancer Prevention Research (Philadelphia)* **8**, *9* (2015), 1–4

[33] M. J. Wensink (2016) Size, Longevity and Cancer: age structure. *Proc. R. Soc. B* **283**, 20161510, 6 pp.

[34] `https://www.cancer.gov/about-cancer/understanding/statistics`

[35] MIT opencourseware, `http://www.rle.mit.edu/rgallager/documents/Poisson.pdf`

[36] `https://seer.cancer.gov`

[37] `https://en.wikipedia.org/wiki/Mortality_rate`

[38] `https://www.cdc.gov/cancer/npcr/uscs/technical_notes/stat_methods/rates.htm`

Dom Eleshnitsa, LTD
e-mail: `maria.beljaeva29@gmail.com`

Received September 1, 2017