Mathematical Biosciences 236 (2012) 16–30

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

New stochastic carcinogenesis model with covariates: An approach involving intracellular barrier mechanisms

Igor Akushevich [a,*], Galina Veremeyeva [b], Julia Kravchenko [c], Svetlana Ukraintseva [a], Konstantin Arbeev [a], Alexander V. Akleyev [b], Anatoly I. Yashin [a]

[a] Center for Population Health and Aging, Duke University, Durham, NC, USA
[b] Urals Research Center for Radiation Medicine, Chelyabinsk, Russia
[c] Duke Cancer Institute, Duke University School of Medicine, Durham, NC, USA

ARTICLE INFO

Article history:
Received 18 October 2010
Received in revised form 4 December 2011
Accepted 9 December 2011
Available online 17 December 2011

Keywords:
Multi-stage models
Cell barrier mechanisms
Stochastic modeling
SEER data
Simulation study

ABSTRACT

In this paper we present a new multiple-pathway stochastic model of carcinogenesis with potential of predicting individual incidence risks on the basis of biomedical measurements. The model incorporates the concept of intracellular barrier mechanisms in which cell malignization occurs due to an inefficient operation of barrier cell mechanisms, such as antioxidant defense, repair systems, and apoptosis. Mathematical formalism combines methodological innovations of mechanistic carcinogenesis models and stochastic process models widely used in studying biodemography of aging and longevity. An advantage of the modeling approach is in the natural combining of two types of measures expressed in terms of model parameters: age-specific hazard rate and means of barrier states. Results of simulation studies allow us to conclude that the model parameters can be estimated in joint analyses of epidemiological data and newly collected data on individual biomolecular measurements of barrier states. Respective experimental designs for such measurements are suggested and discussed. An analytical solution is obtained for the simplest design when only age-specific incidence rates are observed. Detailed comparison with TSCE model reveals advantages of the approach such as the possibility to describe decline in risk at advanced ages, possibilities to describe heterogeneous system of intermediate cells, and perspectives for individual prognoses of cancer risks. Application of the results to fit the SEER data on cancer risks demonstrates a strong predictive power of the model. Further generalizations of the model, opportunities to measure barrier systems, biomedical and mathematical aspects of the new model are discussed.

1. Introduction

The development of new research technologies has resulted in accumulation of large sets of information reflecting some specific features of carcinogenesis at different levels of vital organization. Currently, measurements of dozens, hundreds, and even thousands of indices characterizing the state of biological processes at a molecular level are possible. At the same time, there are currently no reliable methodological ways to apply the knowledge on individual state of an organism to epidemiologic data nor, consequently, to propose a quantitative description of carcinogenesis to the extent of predicting individual risks. In this paper we present a new methodological approach for creating specific models of carcinogenesis which is expected to become one of the first steps in developing carcinogenesis models capable of combining information from individual measurements with many years of broad experience in epidemiologic research. Methodological and substantive backgrounds for the model include: (i) classical multistage mechanistic models of carcinogenesis, (ii) population models of aging and mortality, and (iii) knowledge about specific molecular pathways, both promoting and preventing carcinogenesis at its different stages.

Carcinogenesis modeling has a long history that started with papers of Nordling [1] and Armitage–Doll [2]. These models were suggested to explain the observation that age-specific rates of many common carcinomas increased roughly with the power of age. Then many interesting biological ideas and advanced mathematical methods were developed to understand spontaneous and radiation carcinogenesis. Comprehensive reviews can be found in Refs. [3–7]. A set of these models is a primary methodological background for our approach.

Further progress in modeling carcinogenesis can be achieved by the development of a new formalism capable of including additional information, e.g., measurements of respective risk factors at an individual level or auxiliary information from other sources.

* Corresponding author. Address: Center for Population Health and Aging (CPHA), Duke University, 001 Trent Hall Drive (Box 90408), Durham, NC 27708, USA. Tel.: +1 (919) 668 2715; fax: +1 (919) 684 3861.
E-mail address: igor.akushevich@duke.edu (I. Akushevich).

0025-5564/$ - see front matter © 2012 Published by Elsevier Inc.
doi:10.1016/j.mbs.2011.12.002

Models with the required properties are known in demography and population studies as stochastic process models (SPM), or life tables with covariates. Versions of the models most closely related to the needs of the approach being developed in this paper are described in Refs. [8,9]. Such models allow for inclusion of measurements of covariates (risk factors) and the combined description of their dynamics and survival. Moreover, these models have very useful and well-established mathematical properties. Such population models constitute the secondary part of the methodological background of our approach. The linkage between population and mechanistic – i.e., models of carcinogenesis discussed above – formalisms is provided by the diffusion approximation of the birth–death process described by Tan [10].

The biological background of the approach to carcinogenesis modeling is based on the consideration of carcinogenesis as the dynamic trade-off between two antagonistic forces or processes, promoting or hindering carcinogenesis at its different stages (initiation, promotion, and progression) [11]. Processes promoting the cell malignization are represented by mutations or adverse epigenetic events. Antagonistic processes preventing the neoplastic transformation of the cell and its consequent fixation in the next cell generations are represented by intracellular barrier mechanisms. According to Hanahan and Wienberg [12] there exist up to six intracellular barrier mechanisms (or 'hallmarks of cancer'), which play a determinative role in carcinogenesis from initiation to promotion and progression. These intracellular barrier mechanisms could include self-sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. One more factor, the genetic instability, facilitates the acquisition of other mutations due to defects in DNA repair.

The paper is structured as follows. After a description of methodological background in Sections 2.1 and 2.2, we present the barrier breaking mechanism (BBM) model in Sections 2.3–2.5. The discussion in these sections includes a description of the concept of barrier mechanisms and a mathematical description of the concept from the exact formulation of the model to the parameter estimation procedure for different experimental designs. Several illustrative examples are considered in Section 3 such as the development of the simplest version of the model, its application to the Surveillance, Epidemiology and End Results (SEER) Registry data [13] on solid cancer and leukemia risks, and simulation studies designed to investigate the predictive powers of the model for different experimental designs. Technical details of mathematical development of the model are presented in Appendix A. Discussion of different aspects of the models, including ways for its further generalizations and possible applications, is given in Section 4.

2. Breaking barrier mechanisms (BBM) model

2.1. General population model of carcinogenesis with covariates

A population model capable of describing the dynamics of covariates in connection with the risks of cancer incidence is based on the stochastic process model of human mortality and aging [8,9,14]. This model and its extensions [15–18] employed the Markov property of stochastic process satisfying the diffusion type stochastic differential equation.

Let Y_t be the multidimensional stochastic process with continuous components representing covariates. The following properties are critical for the modeling approach developed in this paper. First, the stochastic process is stopped at a random time associated with the onset of individual's cancer. The current value of Y_t determines the risk of the cancer onset. Second, there are measurements of Y_t for individuals during their follow-up periods. Different approaches to the measurements define different experimental study designs. Third, the dynamics of Y_t is defined by the stochastic differential equation

$$dY_t = A(Y_t, t)\,dt + b(t)\,dW_t, \quad Y_{t_0} \tag{1}$$

Here, $A(Y_t,t)$ is a vector function, $b(t)$ is a matrix of the respective dimension, Y_{t_0} is a random vector of the initial conditions, and W_t is a vector Wiener process with independent components which is independent of the initial value Y_{t_0}. The first term in R.H.S. of Eq. (1) is the deterministic component of this equation that describes the dynamic balance of covariates represented by Y_t, e.g., birth–death dynamics of carcinogenic cells under risk. The stochastic component of the equation describes the effects of an unobservable heterogeneity of cohort members with respect to carcinogenic processes.

Because of random stopping of Y_t associated with cancer incidence, the distribution of Y_t at any time t represents the distribution conditional on surviving, i.e., the distribution of covariates of healthy individuals at any time. Therefore, the hazard rate $\lambda(Y_t,t)$ of random stopping (i.e., cancer onset) at any time is the hazard rate for healthy individuals, and thus, exactly corresponds to the definition of incidence rate. Additional corrections to exclude cancer cases from denominators describing the population under risk typically performed in carcinogenesis modeling, are not required for this approach.

In the general case, the density function for the distribution of Y_t conditional on surviving satisfies the generalized Kolmogorov equations derived by Yashin et al. [8]. If the function $A(Y_t,t)$ is linear with respect to Y_t, and if the order of the dependence of $\lambda(Y_t,t)$ on Y_t is not higher than quadratic, the solution can be obtained in the form of the Gaussian process and the Gaussian property of conditional distribution of the covariate values is guaranteed for any given point in time. This allows a description of carcinogenesis in terms of the two first moments of the multidimensional Gaussian distribution. One advantage of the approach is the existence of the efficient statistical procedure of parameter estimation using data involved with cancer incidence and longitudinally measured covariates.

The described features of the SPM, as well as several other, made this model unavoidable in biodemography, making it useful in modeling of carcinogenesis. These features are worth summarizing. First, covariates are naturally incorporated into the model in addition to the information on individual survival; adding a new covariate does not change the structure of the model. The distribution of covariates conditional on survival is normal if the dynamic stochastic differential equation is linear over covariates and mortality is a quadratic (or linear) function of covariates. Second, the dynamics of covariates is described by the Gaussian stochastic process, characteristics of which (i.e., the vector of means and the variance–covariance matrix) are defined by a system of ordinary differential equations analytically or numerically solvable. Such description is possible due to the conditional Gaussian property of the covariate distribution at any given time which was rigorously proved by Yashin [19]. Third, since the time between surveys may be flexible and the model automatically generates the values of risk factors to fill in missing data, the SPM is the appropriate model for analysis of longitudinal data with irregular measurements. Fourth, there exist versions of the model for different experimental designs, e.g., (i) when only the time of event is measured, (ii) when covariates for individuals are measured, (iii) when covariate dynamics is assigned but they are not measured nor partially measured, and (iv) whole trajectories of covariates are observed. Fifth, there exists an exact procedure of parameter estimation for all experimental designs. Parameter estimates are consistent and identifiable under mild conditions. Finally, there are

straightforward possibilities to include the characteristics of exposure (e.g., dose or dose rate) into the model [20].

Functions $A(Y_t,t)$, $b(t)$, and $\lambda(Y_t,t)$ require further specifications that depend on the type of covariates used. In this paper we assume that the covariates represent some measurable cellular characteristics in the critical tissue (e.g., apoptosis in peripheral blood). We also assume that the respective specifications can be found by considering a mechanistic model of susceptible cell dynamics. Substantive basis for this modeling approach is based on consideration of the concept of barrier breaking as the underlying force of carcinogenesis described recently in details by Veremeyeva et al. [11]. However, before developing the model that is based on the concept of barrier systems and that utilizes the SPM methodology, it is useful to investigate how the described techniques differ from the approaches widely used in multistage carcinogenesis modeling.

2.2. Application of SPM to two-stage carcinogenesis model

Current state of the art of modeling efforts in tumorigenesis relies on a multi-stage hypothesis which is implemented in multistage models of carcinogenesis [21]. The most popular version of the multi-stage models is the Two Stage Clonal Expansion (TSCE) model (Fig. 1). The simplest version of the model, where the number of susceptible normal cells is either constant or described by a deterministic function and where all rates are time-independent, predicts a hazard rate,

$$h(t)=\frac{X(e^{(\gamma+2q)t}-1)}{q(e^{(\gamma+2q)t}+1)+\gamma}, \tag{2}$$

in terms of only three parameters [22]:

$$X=N\mu\mu_1,\quad \gamma=\alpha-\beta-\mu,\quad q=\frac{1}{2}\left(-\gamma+\sqrt{\gamma^2+4\alpha\mu}\right).$$

This is an attractive property of the model. Not all biological parameters (i.e., the number of stem cells N, first μ_1 and second μ mutation rates, and proliferation α and death/differentiation β rates) can be identified using the data on age-specific incidence rates. Recent development of the TSCE model and its applications to individual and population data (including likelihood-based approaches using the TSCE model with time dependent covariates) are described in Section 4.

The sequence within mathematical modeling resulting in Eq. (2) includes the following steps. The starting point is Kolmogorov's equations for probabilities $P_{jk}(t)$ to find exactly j intermediate and k malignant cells at time t (see Ref. [23] Section 3). Then this equation transforms into the partial differential equation for probability generating function $\Psi(y,z;t)=\sum_{jk}y^jz^kP_{jk}(t)$. The function determined for $y=1$ and $z=0$ has a meaning of survival function of cancer risk. The equation for this function in terms of parameters defined in Fig. 1 is

$$\Psi'_t=(y-1)N\mu_1\Psi+(\mu yz+\alpha y^2+\beta-(\alpha+\beta+\mu)y)\Psi'_y. \tag{3}$$

The equation admits an analytical solution (third step, see for example, Refs. [23–25]). The fourth is to calculate the hazard function as $h(t)=-\Psi'_t(1,0;t)/\Psi(1,0;t)$ resulting in Eq (2).

The approach described in Section 2.1 can also be applied to the TSCE model. The stochastic equations for the Wiener-TCSE are $dI_t=(\nu+(\alpha-\beta)I_t)dt+bdW_t$ and $h(t)=\mu I_t$, where I_t is the stochastic variable described by the number of intermediate cells in an organism with no malignant cells. Respective ordinary differential equations for the first and second central moments conditional on survival, read

$$\begin{aligned} m'(t)&=\nu+(\alpha-\beta)m(t)-\mu\gamma(t),\\ \gamma'(t)&=2(\alpha-\beta)\gamma(t)+b^2. \end{aligned} \tag{4}$$

Initial conditions for these equations are $m(0)=m_0$ and $\gamma(0)=0$. These equations can be solved resulting in

$$m(t)=m_0e^{\bar\alpha t}+\frac{\nu}{\bar\alpha}(e^{\bar\alpha t}-1)-\frac{\mu b^2}{2\bar\alpha^2}(e^{\bar\alpha t}-1)^2\gamma(t)=\frac{b^2}{2\bar\alpha^2}(e^{2\bar\alpha t}-1),$$

where $\bar\alpha=\alpha-\beta$. For $b=0$ and $m_0=0$ the model reproduces the predictions of a pure deterministic model (Eq. (13) of Ref. [25]).

The equations for the first and second central moments for TSCE can be obtained directly within classic formalism based on the birth–death process involving the p.g.f $\Psi(y,z;t)$. The moments conditional on the absence of malignant cells (i.e., $k=0$) are obtained by differentiation of $\Psi(y,z;t)$ on y, e.g., $m(t)=\Psi'_y(1,0;t)/\Psi(1,0;t)$. Therefore, the differentiation of Eq. (3) and subsequent algebraic manipulation results in

$$\begin{aligned} m'(t)&=\nu+(\alpha-\beta)m(t)-\mu\gamma(t),\\ \gamma'(t)&=\nu+2(\alpha-\beta)\gamma(t)+(\alpha+\beta)m(t)-\mu\tau(t), \end{aligned} \tag{5}$$

where $\tau(t)$ is the third order central moment conditional on survival (i.e., $k=0$). Eqs. (4) and (5) for the first moment obtained within two formalisms coincide with each other. Furthermore, they coincide exactly with the differential Eq. (14) for $E[Y(t)|Z(t)=0]$ of Moolgavkar et al. [23] ($Y(t)$ and $Z(t)$ are birth–death processes of the numbers of intermediate and malignant cells). Indeed, the first moment $m(t)$ and the second central moments $\gamma(t)$ are interpreted as the mean and variance of the number of injured cells for survivors (i.e., individuals without a malignant cell), i.e., $m(t)$ and $\gamma(t)$ are conditional moments, given the number of malignant cells equals zero. Thus, the quantities $m(t)$ and $\gamma(t)$ exactly correspond to $E[Y(t)|Z(t)=0]$ and $V[Y(t)|Z(t)=0]$ used by Moolgavkar et al. [23].

Comparison of the formulae for the second central moments reveals two distinctions: (i) appearance of the term with $\tau(t)$ in (5) and (ii) difference in some terms which vanishes if $b^2\to\nu+(\alpha+\beta)m(t)$.

The approach described in Section 2.1 resulting in (4) suggests the Gaussian approximation for the stochastic process Y_t. In this approximation all central moments of odd order equal zero, therefore, the term $\mu\tau(t)$ does not appear in (5). The properties of the moments of the Gaussian distribution allow for a closed system of differential equations for moments that can easily be solved analytically or numerically.

The substitution for b^2, required for concurrence of the formulae for the second central moment, is $b^2\to\nu+(\alpha+\beta)m(t)$. This result is in exact accord with the diffusion approximation of the birth–death process developed by Tan [10]. He has formulated his results

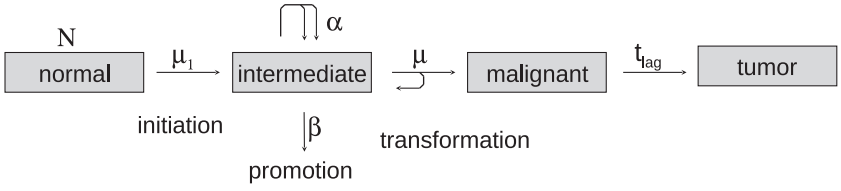

Fig. 1. Two stage clonal expansion model.

in the form of a theorem about the diffusion approximation to the birth–death process defined as $\overline{X}_t = I_t/N$. He has demonstrated that the p.d.f. of $\overline{X}_t$ conditional on its prior value satisfies the Kolmogorov forward equation of a diffusion process with drift and diffusion coefficients equaling $\nu/N + (\alpha - \beta)\bar{x}$ and $(\alpha + \beta)\bar{x}/N$, respectively. Note, that Tan [10] kept only the terms of order $O(N^{-2})$, therefore the term ν appeared in (5), which is of order N^{-2}, and did not appear in his formula for the diffusion coefficient.

At a first glance, the Wiener-TSCE model is the Gaussian approximation of the exact two-stage carcinogenesis. However, this is not accurate because of at least two reasons. The first is the occurrence of dependence on Y_t appeared in the expression for the diffusion coefficient. In the SPM model, the parameter b (and therefore diffusion b^2) is independent of Y_t and this property is important for the SPM model. For example, this property, together with the linear drift and the linear hazard in respect to Y_t, guarantees the Gaussian dynamics for Y_t over time.

Another reason is in the interpretation of the parameter b. In the SPM model the parameter is not simply expressed in terms of the model parameters but reflects the population heterogeneity. Parameters governing the dynamics of Y_t could be distributed in the population. This distribution can be approximated by adding a stochastic variable (sometimes referred to as noise or white noise [26]) which is normally distributed. Because of linearity in respect to these parameters, the stochastic equations described the dynamics of Y_t, and the contributions of the stochastic components that originated from different parameters can be included in the term $b(t)dW_t$.

Thus, the main distinction between classic TSCE and Wiener-TSCE, as well as generally between the two respective formalisms, is in the approach to stochastic components and interpretation of their estimates. The model developed in this paper is based on the Gaussian dynamics. It is specified so that its parameters $A(Y_t,t)$ and $\lambda(Y_t,t)$ are linear over Y_t, and $b(t)$ is independent of Y_t. Therefore, we deal with the linear system of stochastic equations describing the dynamics of Y_t, which is subject to random stopping. Such systems can be solved analytically in many cases, the procedure of maximum likelihood parameter estimation is well-defined for these, and such systems can be easily generalized for describing systems of higher dimensions.

2.3. The concept of barrier-breaking mechanisms of carcinogenesis

The concepts of barrier mechanisms as a foundation for the approach to the mathematical model of carcinogenesis are formulated on the basis of numerous observations collected in multiple studies of intracellular processes, occurring in the norm and in the pathology, including malignant tumors [27]. This concept is based on the following principles. Carcinogenesis represents a set of structural and functional changes in susceptible cells ultimately expressed at a higher level of hierarchy in the tissue of a human body. Detrimental changes at the cell level are caused by an insufficient quality of operation of a complex of mutually interacting barrier mechanisms (e.g., antioxidant defense (AOD), repair systems, apoptosis). The intracellular barrier mechanisms represent the complex of cell responses to events negative for the cell and/or the whole organism. They are combined in a system of cell defense protection from occurrences of genetic damages and their further fixation as mutations, potentially promoting the cell to carcinogenesis. Cell malignization may occur due to an inefficient operation of a part of or of all barrier mechanisms. Hierarchy and complex interaction of barrier mechanisms can compensate for the inefficiency of operation of certain mechanisms by reinforcing others, resulting in a decreased risk of pathology development. Disorders in the barrier mechanism functioning occur as a result of one or more mutations/aberrations or adverse epigenetic events, and there exists a conceptual possibility of measuring the efficiency of each barrier functioning in sensitive cells of an individual.

By adopting the understanding of the barrier mechanisms as dynamic processes hindering carcinogenesis at its different stages (initiation, promotion, conversion), we realize that in each cell there simultaneously exist processes promoting cell malignization (e.g., mutations, epigenetic events) and antagonistic processes preventing the neoplastic transformation of the cell. The latter can include: (i) removal of superfluous free radicals that induce damages in biomolecules, (ii) repair of DNA-damages that occurred through the free radical mechanism or due to other causes, and (iii) apoptosis (in the case if repair is not effective and/or the damage is not compatible with cell functions). What was considered in traditional multi-stage carcinogenesis models as events (i.e., mutational or of combined origin) transferring the cell over stages of initiation, promotion, and conversion, can be considered as results of inefficiency of cell barrier mechanisms within the concept of barrier-breaking mechanisms. A critical conjunction of failure in several mechanisms (e.g., combinations of inefficient repair, violations in apoptosis, activation of telomerase resulting in immortality of the cell) and further appearance of two daughter cells can be considered as its conversion to cancer phenotype.

Thus, the concept of the cell barrier mechanisms of initiation and promotion of carcinogenesis and the model based on this concept naturally generalizes the idea of a multistage model of carcinogenesis. The transfer of a cell from a state to a state might be understood as a failure, or a break, of a certain barrier mechanism in a given cell. Such transfer can occur due to a mutation as in the standard multistage model, due to several mutations, due to adverse epigenetic events or their combinations. A distinguished property of the approach based on this concept is that variables describing the cell state are measurable at the individual level.

As it was discussed in Section 2.1, two components have to be modeled to quantitatively describe time (age)-dependent cancer risk in terms of barrier mechanisms. The first is the dynamics of health states (i.e., the state with different types of intermediate cells and their different amounts) represented by stochastic processes reflecting each of the barrier mechanism and the interaction between them. The second is the model of the risk of solid cancer or leukemia as a function of the health state. If the health state represented by a state of barrier systems is described by a vector of stochastic processes Y_t, then the model is represented by the respective stochastic differential equations and the specific hazard function of risk $h(t,Y_t)$.

The choice of a specific barrier mechanism to be included in the carcinogenesis model depends on the type of the considered cancer, on their rate limiting properties, and also on the possibility of their measurements at an individual level. The methodology developed in this paper is quite universal and can be generalized for a model with a dozen barrier mechanisms, e.g., for all 'hallmarks of cancer' [12]. For simplicity and specificity, we consider a three-component model in which two barriers, say barrier A and barrier B, are considered to be the key components of intracellular barrier mechanisms, and one barrier C plays a promotional role. For example, barriers A and B can be associated with apoptosis and repair systems, respectively, and barrier C can be associated with an antioxidant defense. An important argument for the choice of these mechanisms is the possibilities of barrier-related measurements, e.g., in this example, using cytometry for apoptosis, the comet assay method to measure DNA repair, and methods of analyses of contents of respective proteins for AOD [28].

2.4. Three-component dynamic model of states of barrier mechanisms

The scheme of a compartmental model implementing the sequence of barrier breaking when a cell undergoes changes from a

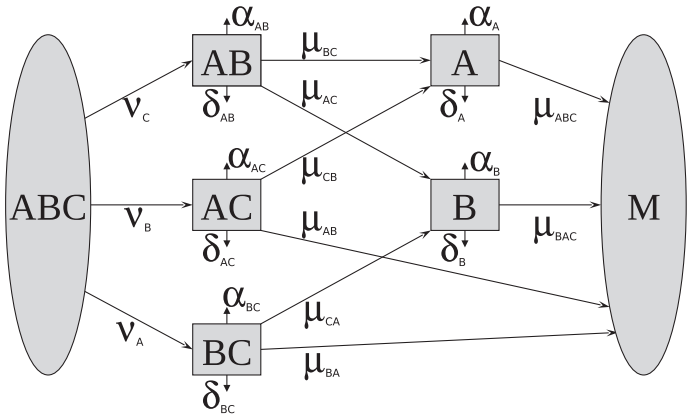

Fig. 2. Scheme of the compartmental model of sequential breaking of barrier mechanisms. Transfers correspond to the failure of a specific barrier.

normal (N or ABC) to a malignant (M) state, is presented in Fig. 2. Different blocks correspond to different cell states, and each cell can be in one, and only one, state. It is assumed in the model, that a cell becomes malignant if barrier mechanisms A and B are broken. Barrier C plays the role of a promoter of the process of malignization, therefore its contribution is non-symmetric in respect to other barrier mechanisms. Letters on the blocks denote which barriers are effective in a certain state. Transfers from a state to a state occur with rates marked by ν or μ at the corresponding arrow. The subscript of ν and the first subscript of μ denote a barrier breaking at the corresponding transfer. The remaining subscripts of μ show the barriers broken before the time of transfer.

Deterministic dynamics of the mean numbers of cells can be described by a system of ordinary differential equations,

$$\begin{aligned}
&m'_N(t) = \varphi(t) m_N(t),\\
&m'_{BC}(t) = (\alpha_{BC} - \delta_{BC}) m_{BC}(t) + \nu_A m_N(t),\\
&m'_{AC}(t) = (\alpha_{AC} - \delta_{AC}) m_{AC}(t) + \nu_B m_N(t),\\
&m'_{AB}(t) = (\alpha_{AB} - \delta_{AB}) m_{AB}(t) + \nu_C m_N(t),\\
&m'_B(t) = (\alpha_B - \delta_B) m_B(t) + \mu_{AC} m_{AB}(t) + \mu_{CA} m_{BC}(t),\\
&m'_A(t) = (\alpha_A - \delta_A) m_A(t) + \mu_{CB} m_{AC}(t) + \mu_{BC} m_{AB}(t),\\
&m'_M(t) = \mu_{AB} m_{AC}(t) + \mu_{BA} m_{BC}(t) + \mu_{ABC} m_A(t) + \mu_{BAC} m_B(t).
\end{aligned} \tag{6}$$

Here, $\varphi(t)$ is the rate of change in the number of normal stem cells often assumed to be zero, e.g., in modeling carcinogenesis in adults. State C is absent in this scheme because this state, i.e., the state with broken barrier A and B, is considered malignant. Therefore, the scheme in Fig. 2, as well as the system of differential equations, is not symmetrical over A, B, and C. Initial values of the system are defined by measurements made at the initial time, or they can be taken from other studies. Note, that we make a usual assumption that the transfer between compartments associated with barrier breaking due to accumulated mutations or adverse epigenetic events, occurs during the cell division. Only the newborn cells become modified, so the number of cells in the original compartment does not change.

The stochastic model is formulated as follows. The state of a healthy individual is modeled by the five-dimensional stochastic process Y_t whose components $Y_{BC}(t)$, $Y_{AC}(t)$, $Y_{AB}(t)$, $Y_B(t)$, and $Y_A(t)$, are associated with the numbers of cells in the intermediate states in Fig. 2. Means of the components satisfy the corresponding equations of system (6). The rates of the newly appearing cells with primary damages $\nu_\cdot m_N(t)$ are considered time-independent. This is the only way for the number of stem cells $m_N(t)$ to appear in the model. The rates of secondary damages and malignant transformation events, $\mu_\cdot$, are also time-independent. Note, assumption on time independence of rates is not essential. The probability of disease development is described by the hazard function, which can be written on the basis of the last equation of the system (6):

$$\begin{aligned}
h(t) = {} & \mu_{AB} Y_{AC}(t - t_0) + \mu_{BA} Y_{BC}(t - t_0) + \mu_{ABC} Y_A(t - t_0)\\
& + \mu_{BAC} Y_B(t - t_0),
\end{aligned} \tag{7}$$

where t_0 is the time period between the first appearing malignant cell and the time of clinical manifestation, i.e., the time of the onset of diagnosis. In what follows we assume that $t_0 = 0$, but its contribution can be easily reconstructed by adding to the list of model parameters and estimating using likelihood (11). How the likelihood function changes for the cases of constant and gamma-distributed t_0 was shown by Meza et al. [29]. Estimates for t_0 for breast cancer and analysis of the correlation effects of this parameter with other model parameters was recently performed by Kravchenko et al. [30].

The system for these five stochastic processes and the hazard function can be written in a matrix form:

$$\begin{aligned}
&dY_t = [a_0 + a_1 Y_t]\, dt + b \cdot dW_t,\\
&h(t) = Y_t^* \tilde{\mu}_1,
\end{aligned} \tag{8}$$

where dW_t is the 5-dimensional Wiener process, a_0 and $\tilde{\mu}_1$ are 5-dimensional vectors, and a_1 and b are 5×5 matrices. The explicit form for Y_t, a_0, a_1, and $\tilde{\mu}_1$, follows from the system (1):

$$Y_t = \begin{pmatrix} Y_{BC}(t)\\ Y_{AC}(t)\\ Y_{AB}(t)\\ Y_B(t)\\ Y_A(t) \end{pmatrix}, \quad a_0 = \begin{pmatrix} \nu_A N\\ \nu_B N\\ \nu_C N\\ 0\\ 0 \end{pmatrix},$$

$$a_1 = \begin{pmatrix} \Delta_{BC} & 0 & 0 & 0 & 0\\ 0 & \Delta_{AC} & 0 & 0 & 0\\ 0 & 0 & \Delta_{AB} & 0 & 0\\ \mu_{CA} & 0 & \mu_{AC} & \Delta_B & 0\\ 0 & \mu_{CB} & \mu_{BC} & 0 & \Delta_A \end{pmatrix}, \quad \tilde{\mu}_1 = \begin{pmatrix} \mu_{BA}\\ \mu_{AB}\\ 0\\ \mu_{BAC}\\ \mu_{ABC} \end{pmatrix},$$

where $\Delta_\cdot = \alpha_\cdot - \delta_\cdot$ for all states. To solve the system means to find deterministic vector function of time for the mean ($m(t)$) and matrix function for the variance ($\gamma(t)$) of the Gaussian stochastic process Y_t. The system of ordinary differential equations for $m(t)$ and $\gamma(t)$ in the matrix form read

$$\begin{aligned}
&dm(t)/dt = a_0 + a_1 m(t) - \gamma(t)\tilde{\mu}_1,\\
&d\gamma(t)/dt = a_1 \gamma^*(t) + \gamma(t) a_1^* + bb^*
\end{aligned} \tag{9}$$

with initial conditions $m(0) = m_0$ (last measurements) and $\gamma(0) = 0$. The hazard rate conditional to survival $\bar{h}(t)$ (i.e., among the surviving population) is

$$\bar{h}(t) = m(t)^* \tilde{\mu}_1, \tag{10}$$

As one can see, the solution is formulated in terms of time dependence of quantities potentially observed in a dataset, i.e., the mean and variances of barrier states, which are modeled by a vector $m(t)$ and a matrix $\gamma(t)$. Solutions of the system (9) model the individual trajectories in the state space until the next measurement.

2.5. Experimental designs and parameter estimation

Five experimental designs might be considered:

Design I. Only the age at onset is observed for individuals of the study cohort.
Design II. The age at onset is detected and auxiliary information, helping to model initial conditions of barrier mechanisms and their dynamics, is used.
Design III. The age at onset is detected and several or all barrier states are measured at regular time intervals.
Design IV. The age at onset is detected and several or all barrier states are measured at irregular time intervals.
Design V. Epidemiological information on the age-specific incidence rate is used in addition to rare measurements of barriers in a randomly selected sub-cohort of a small size.

For any experimental design, the dynamic Eqs. (8) or (9) and (10) define the underlying biological model that is responsible for generating data. In the case of Design I, where only data on age-specific incidence are available, only certain combinations of parameters, but not all biological parameters, are identifiable. However, even in this case the model is capable of evaluating certain features of the underlying carcinogenesis mechanisms. Data collected using Design I are vast and the identifiable combinations of parameters can be assessed with high accuracy. Furthermore, in this case the model admits an analytical solution and, therefore, it can be analytically investigated and compared to the most popular carcinogenesis models such as TSCE.

Designs III and IV represent the typical scenarios when the date at onset is measured in addition to the longitudinal measurements of covariates. If the covariates are measured in regular time intervals (e.g., in surveys), then we are talking about Design III; if they are measured irregularly (e.g., extracted from administrative data) and/or contain missing information (a more realistic situation), then that is Design IV. Since all biological parameters can be estimated using data of such designs, these designs are ideal for application of the model. Currently, the measurements which can be used to inform the model appear episodically (though could become available in the near future as discussed in Section 4), therefore, the approaches capable of linking the information from different types of measurements could be extremely helpful.

Designs II and V provide possibilities for joining data. Design II represents the typical situation in the modeling of biological processes when the information about parameter measurements collected from different human and animal studies are used for partial or complete model estimation (e.g., this type of model estimation was used for hematopoiesis modeling by Colijn and Mackey [31]). Design V combines the information collected in the style of Design I (i.e., large datasets and good accuracy of certain identifiable combinations of parameters) to rare longitudinal measurements of covariates (i.e., collected using Design III and IV). Predictive powers of Design V and Design III are investigated in simulation studies in Section 3.4. The simulation studies proved that only a small part of the data needs to contain information about barrier measurements. This possibility occurs because of assumptions about the same dynamics underlying data on cancer onset and data containing longitudinal measurements of barrier states. It means that if the model parameters are known, the data can be simulated for any of the considered designs or for any combinations of them. In practice, however, smaller and more detailed studies tend to focus on special populations, or may impose eligibility criteria introducing biases. In this case, the respective corrections to model parameters have to be made (e.g., incorporation of radiation dose if cohorts under chronic exposure are considered) or it should be taken into account while developing the biological interpretation of the estimated parameters.

For each design, the model parameters that have to be estimated or modeled are: (i) initiation rates $\nu_{A,B,C}$ and rates of secondary barrier breaking μ's; (ii) all death/differentiation δ's and proliferation α's rates; only their differences contribute; and (iii) initial distribution (means and variances) and the stochastic components b's.

The procedure of parameter estimation is design-specific. For Design I, it is sufficient to calculate age patterns of the incidence rate and fit the model using the least squares. In this case, however, not all model parameters can be identified. We demonstrate it in the next section using a simplified version of the model. In the most general case (i.e., Design IV), parameters can be estimated using the maximum likelihood method [9]:

$$L = \prod_i \bar{h}\left(\tau_i, \widehat{Y}_i(\tau_i)\right)^{\delta_i} \exp\left(-\int_0^{\tau_i} du \bar{h}\left(u, \widehat{Y}_i(u)\right)\right) \times \prod_{j=1}^{k_i} f\left(Y_i(t_j)|\widehat{Y}_i(t_{j-1})\right), \tag{11}$$

where i runs over all individuals in the study cohort; for each individual, the age at onset τ_i is measured (it can be censored; δ_i is the censoring indicator), in addition to each individual k_i measurements of barrier mechanisms $\widehat{Y}_i(t_j)$ performed at times t_j, $j = 1,\ldots,k_i$; $f\left(Y_i(t_j)|\widehat{Y}_i(t_{j-1})\right)$ is the density of the multivariate normal distribution of the states of barrier mechanisms conditional on their last measured values. The vector of means $m(t)$ and the variance–covariance matrix $\gamma(t)$ are the solutions of the systems of ordinary differential equations (9). Such a system has to be solved for each measurement, for each individual, and for each step of the optimization algorithm. The calculation is feasible, though computationally extensive.

3. Illustrative examples

Below, we present four illustrative examples. The first three are devoted to the development of the explicit calculation of the model for Design I, analytical comparison to the TSCE model, and application of the obtained models to a series of data on cancer incidence in the US. The fourth includes simulation studies of Designs III and V.

3.1. A version of the BBM model for Design I

A series of reasonable assumptions can be made to essentially reduce the number of model parameters if the model is applied to data collected using Design I. First, assume that all rates of barrier breaking in normal cells are equal, i.e., $\nu_A = \nu_B = \nu_C = \nu_0$, since genetic and epigenetic events in genes regulating these systems are similar. Second, assume that the proliferation-apoptosis balance is described by the differences (e.g., $\Delta_{BC} = \alpha_{BC} - \delta_{BC}$) and barrier C does not essentially impact this balance, so it can be neglected and therefore, $\Delta_{AB} = 0$, $\Delta_{AC} = \Delta_A$, and $\Delta_{BC} = \Delta_B$. Also, make an additional simplifying assumption that $\Delta_A = \Delta_B = \Delta \geqslant 0$. Finally, assume that the rates of barrier breaking are proportional to the initiation rate with an amplifying coefficient dependent on the state, i.e., $\mu_{CA} = \mu_{BA} = \nu_0 C_A$, $\mu_{CB} = \mu_{AB} = \nu_0 C_B$, $\mu_{AC} = \mu_{BC} = \nu_0 C_C$,

$\mu_{ABC} = v_0 C_B C_C$, $\mu_{BAC} = v_0 C_A C_C$. In addition, assume that $C_A = C_B = C$. Four free parameters (v_0, Δ, C, C_C) for the description of transition rates remain in the model after these assumptions.

The system (9) and (10) for $m(t)$ and $\gamma(t)$ can be solved analytically in the general case, however, in such a four-parameter approximation, each step in the solution can be illustrated by compact intermediate solutions. Details of this calculation are given in Appendix A. The explicit expression for the conditional hazard obtained without any additional assumptions reads

$$\bar{h}(t) = \bar{h}_0(t) + \bar{h}_N(t) + \bar{h}_b(t), \tag{12}$$

where

$$\bar{h}_0(t) = \Delta(\varepsilon_1 E(1+\varepsilon\tau)(m_{10}+m_{20}) + \varepsilon E(m_{40}+m_{50}) + 2\varepsilon\varepsilon_2(E-1)m_{30}),$$

$$\bar{h}_\phi(t) = 2\phi(\varepsilon_1(E-1) + \varepsilon\tau(\varepsilon_1 E - \varepsilon_2) - \varepsilon(\varepsilon_1 - \varepsilon_2)(E-1)),$$

$$\bar{h}_b(t) = -\frac{1}{2}\varepsilon_1^2(E-1+\varepsilon(\tau E - E + 1))^2(b_1+b_2) - \frac{1}{2}\varepsilon^2(E-1)^2(b_4+b_5) - 2\varepsilon^2\varepsilon_2^2(\tau - E + 1)^2 b_3.$$

The following notation is used to simplify the expressions:

$$\phi = v_0 X_N, \quad \tau = t\Delta, \quad \varepsilon = \frac{v_0 C C_c}{\Delta}, \quad \varepsilon_1 = \frac{v_0 C}{\Delta} = \frac{\varepsilon}{C_c},$$
$$\varepsilon_2 = \frac{v_0 C_c}{\Delta} = \frac{\varepsilon}{C}, \quad E = \exp(t\Delta). \tag{13}$$

Since v_0 represents the rate of mutations, which are rare events, the dimensionless parameters ε, ε_1, and ε_2 are small quantities and can be used for the expansion of $\bar{h}(t)$ into a respective series. As in many other mechanistic models, ϕ is interpreted as the rate of creating the intermediate cells, and because of factor X_N this parameter is not considered small.

The first term in Eq. (12), $\bar{h}_0(t)$, describes the contribution of the initial state: each term contributed to $\bar{h}_0(t)$ is proportional to one of the barrier state means taken at the initial time. If one assumes that initial risk is zero, then this term does not contribute to the total hazard. The leading term in expansion over ε's can be represented as

$$\bar{h}_0(t) = \bar{h}_0(0) + (E-1)\bar{h}_0(0) + o(\varepsilon),$$

where $\bar{h}_0(0) = \Delta(\varepsilon_1(m_{10}+m_{20}) + \varepsilon(m_{40}+m_{50}))$.

The second term, $\bar{h}_\phi(t)$, is the main contribution which is an analog of the TSCE model given by Eq. (2). If only to keep the leading contributions of the hazard rate (2) and $\bar{h}_\phi(t)$ in the limit of small mutation rate, we will have the exact same results:

$$h_{TSCE}(t) = \frac{X_N \mu\mu_1}{\alpha - \beta}(e^{(\alpha-\beta)t} - 1) + o(\mu),$$

$$\bar{h}_\phi(t) = 2\phi\varepsilon_1(E-1) + o(\varepsilon) = \Delta^{-1}(v_A + v_B)X_N\tilde{\mu}(e^{\Delta t} - 1) + o(\tilde{\mu}),$$

where $\tilde{\mu} = \mu_{AB} = \mu_{BA} = \Delta\varepsilon_1$. Since $\Delta \equiv \alpha - \beta$ and $\mu_1 = v_A + v_B$, both expressions are identical. Note that there are only two pathways from normal to malignant cells given in Fig. 2, which do not include the events of breaking barrier C, and contribute to $\bar{h}_\phi(t)$ in this approximation.

The third term in (12) reflects population heterogeneity in the change of the barrier states. This term contains only contributions proportional to constants b's, which can be not small. The leading term in expansion over ε's is of the second order and negative for all t:

$$\bar{h}_b(t) \approx -\frac{\varepsilon_1^2}{2}(E-1)^2\Big(b_1 + b_2 + C_c^2(b_4+b_5)\Big).$$

Thus, the minimal model included only the leading terms from all three terms in (12) is

$$\bar{h}_m(t) = p_0 + p_1(e^{t\Delta} - 1) - p_2(e^{t\Delta} - 1)^2. \tag{14}$$

This model has four parameters for fitting, i.e., Δ and p_0, p_1, p_2. First term p_0 characterizes the hazard rate in the time of cohort forming. Each of the model parameters is restricted to being positive. The model with all possible contributions has six positively-definite coefficients

$$\bar{h}(t) = p_1 e^{t\Delta} + p_2\Delta t e^{t\Delta} - p_3\Delta t + p_4(e^{t\Delta} - 1) - p_5(e^{t\Delta} - 1)^2.$$

In practical data analysis, several parameters can be correlated (especially at small Δ). That is why we use another reduced form of the model:

$$\bar{h}_r(t) = p_0 + p_2\Delta t(e^{t\Delta} - 1) + p_3\Delta t - p_5(e^{t\Delta} - 1)^2. \tag{15}$$

3.2. TSCE and BBM models

Typical plots of the shape provided by TSCE and the two versions of BBM model given by Eqs. (14) and (15) are presented in Fig. 3. At small values of time after the age of the cohort forming, the TSCE predicts a slow but accelerated increase of the risk with age. Then the speed of this increase slows down, the curve undergoes the inflection and leveling off. The inflection is given for age

$$a_f = \frac{1}{\gamma + 2q}\log\frac{\gamma - q}{q} = \frac{1}{\gamma}\log\frac{\gamma^2}{\alpha\mu} + o(\mu).$$

The plateau appears at the level of

$$h_{\max} = \frac{X}{q} = \frac{N\mu_1\gamma}{\alpha} + o(\mu).$$

The TSCE curve was calculated for $X = 10^{-7}$ year^{-2}, $q = 0.625 \cdot 10^{-5}$ year^{-1}, and $\gamma = 0.447$ year^{-1}. For these values $a_f = 25$ years and $h_{\max} = 1.6 \cdot 10^{-2}$ years^{-1}. Note that the inflection point is essentially defined by γ, the parameter reflecting proliferation/apoptosis balance in the intermediate cells, and the parameter has to be large (e.g., in comparison to respective prediction given by the BBM model as discussed below) to predict the inflection point at the reasonable age.

The BBM model has three contributions, one of them, $\bar{h}_\phi(t)$, is analogous to h_{TSCE}, and two others reflect the new features

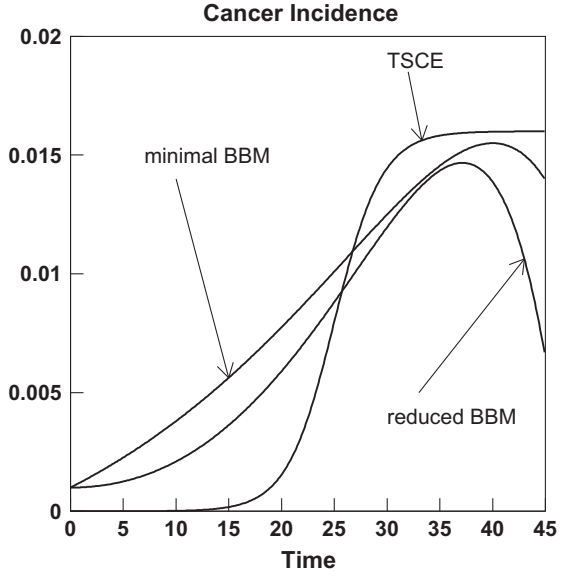

Fig. 3. Typical shape of age dependence of cancer risk provided by the TSCE model (a) and two versions of BBM models, i.e., the minimal and reduced models given by Eqs. (14) and (15).

captured by the model. They are a non-zero initial hazard and the noted stochastic component allowing for description of deceleration of the hazard rate occurred for different cancer age patterns. Fig. 3 shows two curves of the BBM model presented by Eqs. (14) and (15). These two models differ by the type of approximation of $\bar{h}_\phi(t)$, while the initial hazard represented by p_0 and the term with p_5, which is responsible for deceleration of the hazard rate at advanced ages, are identical in both models. The hazard rate of the minimal model can be rewritten in terms of parameters characterizing the geometry of the shape, initial hazard p_0 and the point of maximum $(t_{\max}, \bar{h}_{\max})$:

$$\bar{h}_m(t) = \bar{h}_{\max} - (\bar{h}_{\max} - p_0)\left[\frac{e^{\Delta t_{\max}} - e^{\Delta t}}{e^{\Delta t_{\max}} - 1}\right]^2.$$

The curve corresponding to the minimal model in Fig. 3 is calculated for $p_0 = 0.001\ \text{year}^{-1}$, $t_{\max} = 40$ years, $\bar{h}_{\max} = 0.0155\ \text{year}^{-1}$, and $\Delta = 0.05\ \text{year}^{-1}$. The curve for reduced model is calculated with the same p_0 and Δ as well as for $p_2 = 0.005\ \text{year}^{-1}$ and $p_5 = 0.00125\ \text{year}^{-1}$.

3.3. Application to the SEER Registry data

This model can be used for analyses of age-specific incidence rates as observed in the SEER Registry data, which has been in existence since 1973 and currently covers 26% of the US population [13]. Fifteen cancers were selected for analyses of their age patterns using the model. The non-linear least squares implemented in SAS Proc NLP were used for parameter estimation. The results of the fits are presented in Table 1 and Fig. 4. Two models were used for analyses of each age pattern. The first is the 4-parameter minimal model given by Eq. (14), and the second is the 5-parameter reduced model given by Eq. (15) (Table 1 provides the information on parameter estimation and model selection). Fig. 4 demonstrates the results of the fits of age-specific cancer rates (per 10 000 of population). Rates and model predictions for different cancers are rescaled to use the same scale on all plots and to compare them for different cancers. The real rate for a specific cancer is calculated by division of values obtained from the plot to corresponding rescaled factor referred at each plot. Quality of the fit ($\chi^2/d.o.f \approx 1$) is good for the majority of age patterns. Exceptions include lung and breast cancers (for which the description is still satisfactory), and cancers of prostate and corpus uteri (for which the description is poor).

The age pattern of the lymphoid leukemia incidence rate (the ICD-9-CM code 204, excluding 204.1) has a sharp peak at ages 2–5 years old. The description of the whole age region of the incidence pattern, including this peak, within the same model is challenging. Using the BBM model, such a description is possible by introducing a mixture of two sub-cohorts with different distributions of initial (e.g., genetic) damage. The sub-cohort with a strong genetic predisposition to the disease is responsible for this peak. The mixture model is constructed as

$$\bar{h}(t) = \frac{wS_1(t)}{S(t)}\bar{h}_1(t) + \frac{(1-w)S_2(t)}{S(t)}\bar{h}_2(t); \quad S(t) = wS_1(t) + (1-w)S_2(t);$$

where $\bar{h}_1(t)$ and $\bar{h}_2(t)$ are hazard functions (12), $S_1(t)$ and $S_2(t)$ are corresponding survival functions, and w is the parameter describing the initial proportion between the two cohorts. Altogether, the mixture model includes 13 parameters, i.e., 6 for each hazard and w. Results of the fits are presented in Fig. 5. $\chi^2/d.o.f$ for men and women are 1.89 and 1.63, respectively. For males, the estimated Δ's for two sub-cohorts are $\Delta_1 = 0.0425\ \text{years}^{-1}$ (sub-cohort with genetic predisposition), $\Delta_2 = 0.0041\ \text{years}^{-1}$ (sub-cohort without genetic predisposition) and weight of the first sub-cohort is $w = 0.0001$. The results are similar for females: $\Delta_1 = 0.0621\ \text{years}^{-1}$, $\Delta_2 = 0.0041\ \text{years}^{-1}$, and $w = 0.0001$. Further improvement in description of this age pattern is possible by modeling the distribution of initial damages by a continuous distribution rather than by a mixture of two sub-cohorts.

There are two principal reasons why the quality of fit for several cancers is not ideal. The first is the hidden heterogeneity in data due to different stages at cancer diagnosis, histotypes, genetic predisposition, contributions of environmental risk factors, and period-cohort effects. Second, the considered model has a series of simplifying assumptions, and several of them are strong. One such assumption is the equality $\Delta_A = \Delta_B = \Delta \geqslant 0$. If barrier A is

Table 1
Parameter estimates of minimal and reduced models *of SEER incidence rates*.

Cancer	Sex	p_1	p_2	p_3	p_4	p_5	Δ
Lung	M	0.000117	0.00490	0.000040		0.0016376	0.038
Lung	F	0.000091	0.00530	0.000193		0.0023564	0.030
Breast	F	0.000997			0.00447	0.0014786	0.024
Prostate	M	0.000045	0.00450	−0.000632		0.0011051	0.047
Esophagus	M	0.000013	0.01950	0.000045		0.0142575	0.011
Esophagus	F	0.000003	0.00530	0.000008		0.0037852	0.010
Stomach	M	0.000028	0.00030	0.000059		0.0000750	0.040
Stomach	F	0.000020	0.00000	0.000025		0.0000044	0.050
Colon	M	0.000071	0.00070	0.000076		0.0001196	0.049
Colon	F	0.000064	0.00030	0.000164		0.0000453	0.051
Rectum	M	0.000029	0.00320	0.000122		0.0016440	0.021
Rectum	F	0.000022	0.00110	0.000156		0.0004753	0.021
Ovary	F	0.000076			0.00053	0.0001424	0.027
Corpus uteri	F	0.000073	0.06540	0.002077		0.0531417	0.010
Pancreas	M	0.000020	0.00040	0.000069		0.0001232	0.038
Pancreas	F	0.000014	0.00020	0.000033		0.0000379	0.044
Liver	M	0.000017			0.00012	0.0000105	0.046
Liver	F	0.000005			0.00002	0.0000004	0.073
Brain	M	0.000045			0.00006	0.0000041	0.056
Brain	F	0.000030			0.00003	0.0000019	0.061
Kidney	M	0.000040			0.00021	0.0000175	0.050
Kidney	F	0.000028			0.00008	0.0000055	0.055
CLL	M	0.000007	0.00450	0.000054		0.0023555	0.011
CLL	F	0.000003	0.00040	0.000012		0.0001280	0.022
Myeloid leuk.	M	0.000026			0.00001	0.0000001	0.100
Myeloid leuk.	F	0.000023			0.00001	0.0000002	0.075

All parameters are in 1/year.

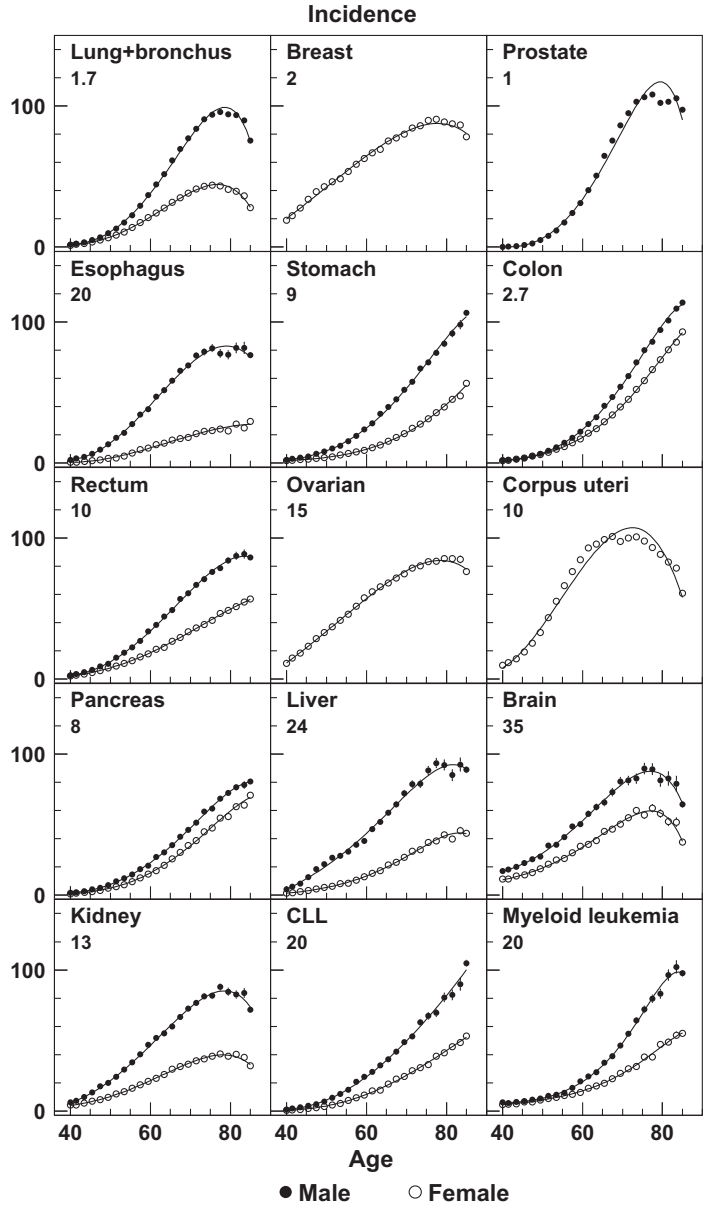

Fig. 4. Age-patterns of cancer incidence rates (dots and standard error bars) calculated using SEER data and BBM model prediction (solid lines) for age patterns of 15 cancer sites.

interpreted as apoptosis, then Δ_A is positive but it can differ if barrier B is interpreted as repair (for which death/differentiation rate can be even larger than proliferation). The potential for further improvement includes: (i) the use of a model with non-equal parameters for barriers A and B with restrictions (e.g., $\Delta_A > \Delta_B$) to make this model identifiable, (ii) the use of additional information on age-specific distribution of barrier mechanisms, i.e., employing Design II, and (iii) performing prospective measurements of individual barrier states, i.e., employing Design III or Design IV.

3.4. Simulation studies

Two scenarios are considered in simulation studies. The first deals with Design III, and the second deals with Design V. In both cases, the model parameters to be estimated are (i) initiation rates (ν_A, ν_B, ν_C), (ii) rates of promotion and conversion ($\mu_{AB,BA,AC,CA,BC,CB,ABC,BAC}$), and (iii) differences between proliferation and apoptosis/differentiation rates ($\Delta_{A,B,C,AB,AC,CA}$). True values are taken as in the simplified model described in Section 3.1. The results of simulation studies are presented in Table 2: one hundred datasets with 100 000 person years each are simulated assuming that covariates Y_t are annually measured.

In the first study (that corresponds to Design III), all simulated data were used for parameter estimation maximizing the likelihood (11). Then parameters estimated for each simulated dataset were averaged over all simulated datasets resulting in the values of the mean and standard errors (SE) and compared with true values. The quantity RAT estimated as $RAT = (mean\text{-}true)/SE$ is the characteristic of the quality of reconstruction of corresponding variables and related to the p-value of the acceptance of the hypothesis that the mean obtained by averaging over parameter estimates from all simulated databases coincides with the true values used for simulations.

Design V, in which data are combined from epidemiological measurements (as in Design I) and physiological (i.e., barrier state) measurements would be the most promising and beneficial for

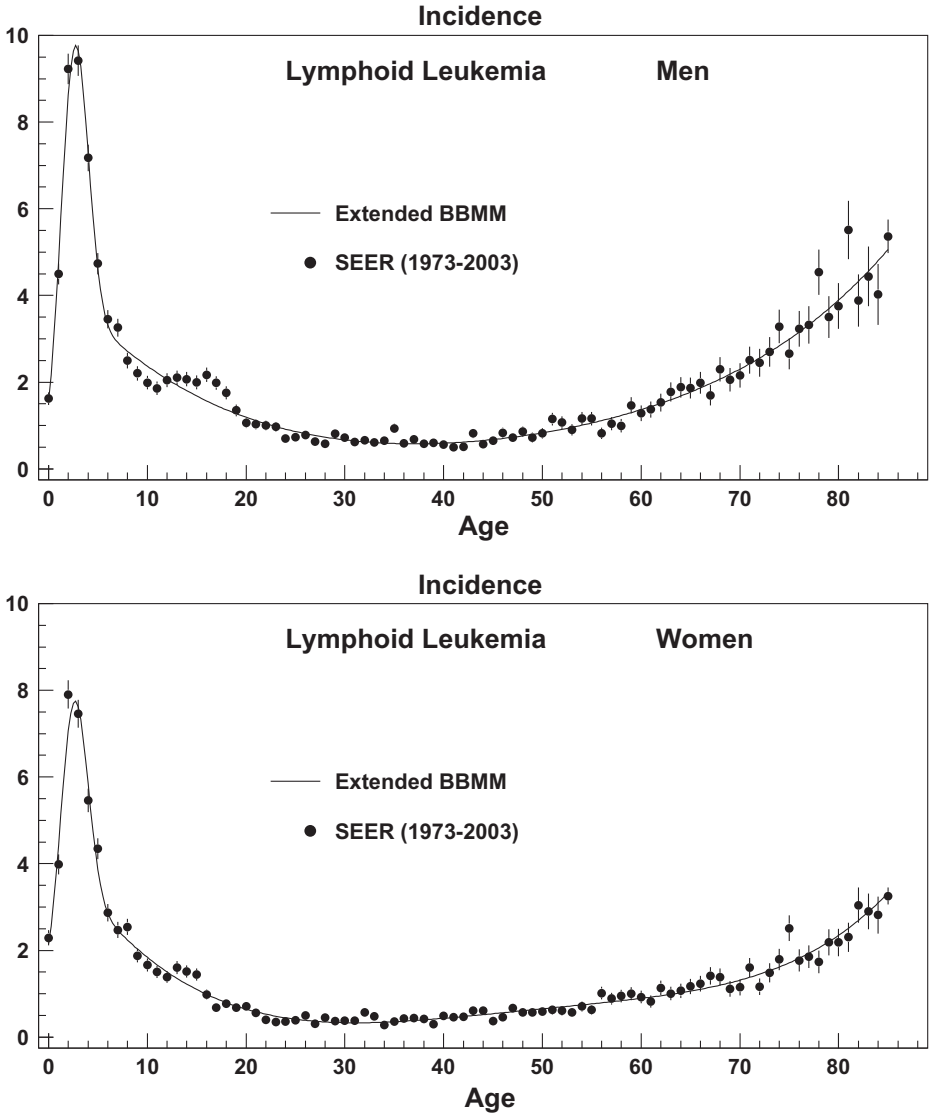

Fig. 5. Lymphoid Leukemia incidence rate (dots and error bars) and predictions by the extended BBM model (solid lines).

Table 2
Results of simulation studies for Designs III and V (marked by asterisk).

	$v_A\ 10^{-6}$	$v_B\ 10^{-6}$	$v_C\ 10^{-6}$	$\Delta_{BC}\ 10^{-3}$	$\Delta_{AC}\ 10^{-3}$	$\Delta_{AB}\ 10^{-7}$	$\Delta_B\ 10^{-3}$	$\Delta_A\ 10^{-3}$	$\mu_{CA}\ 10^{-5}$	$\mu_{BA}\ 10^{-5}$	$\mu_{CB}\ 10^{-5}$	$\mu_{AB}\ 10^{-5}$	$\mu_{AC}\ 10^{-6}$	$\mu_{BC}\ 10^{-6}$	$\mu_{ABC}\ 10^{-4}$	$\mu_{BAC}\ 10^{-4}$
True	1.0	1.0	1.0	1.0	1.0	0	1.0	1.0	2.0	2.0	2.0	2.0	5.0	5.0	1.0	1.0
Mean	1.0	1.0	1.0	1.0	1.0	−1.1	0.9	0.9	2.0	2.0	2.1	2.0	5.2	4.0	2.0	−0.6
SE	<0.1	<0.1	<0.1	<0.1	<0.1	2.2	0.1	<0.1	0.1	0.1	0.2	0.1	0.8	2.3	0.8	0.6
RAT	−0.4	0.9	1.1	−0.4	−1.4	−0.5	−1.3	−1.3	0.1	0.1	0.6	0.2	0.2	−0.5	1.2	−2.7
RAT*	−1.3	0.4	2.3	0.8	−0.5	−2.2	−3	−2.5	−1.5	0.1	0.7	0.2	2.0	−0.6	1.2	−2.7
SE*/SE	10	10.3	10.6	10.7	9.9	11.1	11.9	11.0	9.8	1.0	10.8	1.0	10.1	10.8	1.0	1.0

Initial true values of the parameters are taken for the 4-parameter model; all rates are in 1/year.

efficient data collection and parameter estimation. To illustrate that fact, we estimated the parameters assuming that information on dynamics of barrier state variables was available only for 1% of randomly selected individuals. The likelihood (11) is generalized as

$$L = \prod_i \bar{h}\left(\tau_i, \widehat{Y}_i(\tau_i)\right)^{\delta_i} \exp\left(-\int_{\tau_i}^{0} du \bar{h}\left(u, \widehat{Y}_i(u)\right)\right) \times \left[\prod_{j=1}^{k_i} f\left(Y_i(t_j)|\widehat{Y}_i(t_{j-1})\right)\right]^{\bar{\delta}_i},$$

The novelty in this formula comparing to (11) is the indicator $\bar{\delta}_i$ that shows that the measurements of X_t are available for the individual i. There are parameters ($\mu_{AB,BA,ABC,BAC}$) which are not involved in $f\left(Y_i(t_j)|\widehat{Y}_i(t_{j-1})\right)$, therefore they will be equally estimated for Designs III and V. This fact is illustrated by the ratio of statistical errors in the last row of Table 2: the ratios for the parameters are equal to one.

The conclusion from the simulation studies are that the model parameters are identifiable if dynamics is measured for all or a part of individuals. Dynamic parameters (e.g., all v's and all Δ's) are

defined quite well; while identification of hazard parameters, i.e., μ_{BA}, μ_{AB}, μ_{BAC}, μ_{ABC}, requires large statistics, especially for rare events. However, since these hazard parameters are not involved in the part of likelihood which is responsible for description of dynamics, Design V would be beneficial. Thus, the results of the simulation studies show that Design V for the model estimation is the most perspective because model parameters for hazard require much more data for their estimation but respective datasets are extensive (e.g., SEER data), while parameters for dynamics are rarely measured, but they are relatively well-estimable. The approach opens broad possibilities for combining data sets. The only underlying assumption is the same underlying model generating data of different designs.

4. Discussion

Biologically-motivated mathematical models of carcinogenesis constitute a complementary approach to empirical analyses and standard statistical models capable of enriching empirical findings by new modeling-associated results and, therefore, are important for understanding carcinogenesis mechanisms. The two-stage carcinogenesis model [32] and its extensions have been used for analysis of experimental data obtained from the rodent models [33–36] and from epidemiologic datasets of lung and breast cancers [37–39], colon [40], and solid cancers and leukemia in the Life Span Study (LSS) cohort [41]. In the recent decade, the approaches to carcinogenesis modeling have been applied to the analyses of the SEER registry data such as the incidence data on colon/colorectal cancer [42–44], colon and pancreatic cancers [45], pleural and peritoneal mesothelioma [46], acute lymphoblastic leukemia in children [47], breast ductal and lobular carcinomas [30], and multiple pathways of colon cancer [48]. In this study, we extended the possibilities of such analyses by introducing the potentially measurable covariates into the model. This approach creates a 'bridge' between the analyses of population-based datasets linked to individual longitudinal measurements of specific cancer-related characteristics such as apoptosis and oxidative stress. The BBM model relies on the biological concept of barrier mechanisms in a cell by playing a key role in preventing the cell malignant transformation. The dynamics of covariates is induced by the dynamics of the intracellular barrier breaking, thus, the model connects three levels: (i) a cellular level at which the dynamics of the processes is determined, (ii) an overall organism level at which covariates are measured, (iii) a population level at which such characteristics as disease incidence and mortality are predicted. An advantage of the BBM model is that cell dynamics is described in terms of quantities which are not only important characteristics of carcinogenesis but potentially measurable at the cellular level. The model simultaneously describes the stochastic dynamics of covariates and hazard function. Incorporation of covariate dynamics requires neither new model parameters nor additional assumptions, but is defined by the structure of the BBM model. The covariate dynamics is predicted by the model and parameters describing the dynamics are estimated in addition to the parameters describing hazard function. Since both groups of parameters are related (e.g., assuming the same rate of barrier breaking in all states in Fig. 2), this approach allows to improve the parameter estimates. This property distinguishes our model from other approaches where hazard function is related to covariate measurements [29,49].

Formally, the model is presented by a stochastic process with random stopping associated with the onset of cancer. Since the process satisfies the system of stochastic differential equations which is linear in respect to covariates (i.e., variables describing the numbers of cells in states) and initial values are described by a normally distributed random vector, then the solution of the system is a multivariate Gaussian Markov process. Using a normal multivariate stochastic process for description of cell dynamic allows us to apply all attractive features of the well-developed formalism of Markov processes, providing a great flexibility in specification and further generalizations of the model. At a first glance, the stochastic process models (and BBM model in particular) are models with a philosophy different than the multistage carcinogenesis models and TSCE, in particular. However, in spite of the TSCE dealing with a Poisson process driving a birth–death process and the dynamics of the BBM model being Gaussian, there are many common features in both types of approaches. We demonstrated in Section 2.2 that dynamics for first two moments generated by the two models formally coincide under mild conditions. Background for the similarity of both approaches is the diffusion approximation of the birth–death process [50,51], specified by Tan [10] for the two-stage carcinogenesis model. However, the accordance is not complete, and this can be used for modeling further model development for multidimensional systems. One opportunity is to involve higher order approximations based on evaluating third and higher order moments or cumulants. Further extension might require the use the rich mathematical formalism of non-linear differential equation in Ito and/or Stratonovich versions to further specify properties of the base stochastic process [51].

The approach developed in this paper allowed to reflect on several important properties of carcinogenesis. The first is the description of the different types of intermediate cells. One key quantity in carcinogenesis models is the difference between proliferation and apoptosis rates, and this quantity has to have different estimates (and, maybe, even different signs) for intermediate cells with broken apoptosis and broken repair. Another feature of the approach based on BBM model is that it is capable of taking into account population heterogeneity and describing the decline in incidence rates observed at advanced ages for many cancers. One more advantage is in the BBM model providing cancer risk estimates conditional on individual health history. This approaches us to possibility of individualized medicine.

Thus, the suggested approach combines recent methodological and substantive developments in modeling cancer risk (including the multiple pathway models, the concept of cancer hallmarks) and provides new techniques for description of multistage and multiple pathways models of carcinogenesis with the measured covariates. The multi-pathway nature of our model is not new itself, however the biological content allowing us to define the model in terms of cancer-hallmark measures and the mathematical structure allowing for consistent parameter estimates for several experimental designs involving measures of cancer hallmarks, and relating them to cancer risk, is innovative. The BBM is a complex model involving more than twenty parameters. This number of parameters is large, but this does not result in problems because the parameters have clear biological sense and can be identified given the longitudinal measurements of covariates represented the states of barrier systems.

4.1. Biomedical aspects of BBM model

The concept of barrier mechanisms could potentially serve as a bridge between epidemiologic and mechanistic description of carcinogenesis through inclusion of molecular biology findings [52,53]. The approach based on this concept could allow for inclusion into the model not only of a single gene alteration (which reflects only a small component of each tumor's mutational composition), but also of combined contributions of multiple genes to carcinogenesis (which could represent the same barrier and/or be functionally equivalent – i.e., affecting net cell growth through the same molecular pathway) [54]. Since barrier breaking just

represents such combined events, this approach provides with opportunity to focus on cancer pathways rather than on individual genes when studying carcinogenesis [55]. This consideration is in agreement with the results obtained from molecular studies, e.g., the common features of the general genomic 'landscapes' described for both breast and colorectal tumors [56].

Recent studies demonstrated that the barriers and, therefore, the models of barrier statuses, may have clinical implications: cancers of different histotypes originated from the same organ – such as adenocarcinomas and squamous cell carcinomas – could differ with the involvement of barriers in their carcinogenesis. For example, while the inhibition of apoptosis plays a more important role in adenocarcinomas of cervix uteri, for cervical squamous cell carcinomas the tumor-invasion related factors could be more important [57]. Therefore, the choice of barriers to be included in the carcinogenesis model could be cancer-site and/or histology specific. Thus, known properties of a considered cancer can be used for model specification, which is advantageous when experimental information is limited. Similarly, the different roles various barriers could play in carcinogenesis could also be risk factor specific. That could allow for speculations about the associations with specific cancer risk factors, which potentially could 'act' through different barriers thus affecting the susceptible populations.

4.2. Possibility to measure the state of barrier systems

However, to be able to incorporate barrier mechanisms in human carcinogenesis model, the critical question is the possibility to measure barrier systems at an individual level. The modern state of science opens the broad possibilities for individual measurements of states of barrier mechanisms. For estimating the probability of BBM, the standard loading tests (e.g., experiments with additional exposure and measurements of barrier mechanism response to newly created damages) are useful. A genotoxic exposure entails an additional damage to the genome which should either be repaired or eliminated. A comparison of the values studied in the sample, tested before and after loading, will enable us to draw a conclusion of the effective functioning of the protective mechanisms. A realizable scheme of using loading tests for estimating fractions of failure of specific barrier mechanisms are sufficiently simple and accessible for implementation. The state of barrier systems can be sensitive to the past medical history and/or exposure to risk factors. For example, among exposed to IR individuals, those who manifested leucopenia and/or had chronic radiation syndrome (CRS) during the early after-exposure period had lower Cu/Zn-SOD concentration, a significantly increased concentration of nitric oxide, and a greater apoptotic frequency in peripheral blood lymphocytes compared to exposed individuals without leucopenia and CRS. The persistence of chromosome aberrations and somatic mutations in the CRS cohort is indicative of an exhaustion of the anti-oxidative stress mechanisms responding for so many years after the exposure, leading to genomic instability [53].

Other promising approaches of measurements and modeling barrier states are based on actively developing technologies of biochips providing estimation of gene expression and SNPs analysis. Probability of barrier breaking is associated with the fraction of mutations in gene comprising respective genetic pathways. Expression of allele genes could be affected in genes involved in barrier mechanisms: analysis of such damages in different clusters of functionally related genes could allow not only to estimate how effective the current barrier function is, but, probably, to make a prognosis on the probability of its dysfunction with time (which is important for modeling). Application of modern technologies would allow us to extend the number of measured model parameters with simultaneous broad coverage of the possible mechanisms underlying carcinogenesis. The developed model can be specified to deal with gene expression data by (i) extending the dimensionality of Y_t in the base model to include all (or the most essential part) of gene expression scores, (ii) creating scores describing specific genetic pathways and considering these scores as covariates Y_t, and (iii) categorizing the measured values of gene expression and applying the methods of latent structure analyses (e.g., linear latent structure analysis [58,59]) to identify scores which could serve covariates Y_t.

4.3. BBM approach and Markov models

The covariates in BBM model reflect the health state of an individual (i.e., the states with different types of intermediate cells and their different amounts) and define the probability of stopping time of the stochastic process associated with cancer onset. The made assumptions about susceptible cell dynamics and the probability of stopping time result in the linear system of equations that admits solution in the form of a multivariate Gaussian process. Thus, the BBM model is based on the Gaussian approximation of the general Markov model. The dynamics of covariates and covariate-dependent hazard functions are linked and described in the general Markov model using the Kolmogorov–Fokker–Planck equation for probability distributions of the covariates. Examples of such models are given by Yashin et al. [8,60]. Boundary conditions in such models can be used to define the absorbing states and to avoid the appearance of non-negative values of covariates.

As demonstrated by Tan [10], the diffusion approximation to the classic multistage cancer model based on the birth–death process can be developed under the mild assumptions resulting in a version of the Markov model in which the fraction of injured cells plays the role of a covariate. Therefore, the model corresponds well to our approach. In the general Markov model, the diffusion coefficient b can depend on a covariate and it is proportional to $\sqrt{X(t)}$ in Tan's model. The stochastic differential equation defines the continuous stochastic process, therefore $X(t)$ cannot become negative for positive drift.

The general BBM model (without the Gaussian approximation) is complicated and requires extensive calculations (involving the solution of the partial differential equations for each step of the optimization procedure). Tan's model can be solved using the Laguerre polynomials, however, such solution is also complicated for practical tasks. For example, it is not clear (i) how to implement information about covariates measured longitudinally, (ii) how to develop a likelihood-based scheme for parameter estimation involving covariate measurements and risks of cancer simultaneously, and (iii) how to visualize the relation (i.e., common features and distinctions) to the classic cancer models, e.g., TSCE. The present paper was focused on these issues. The approximate solution in the form of Gaussian process used in the present paper is a reasonable compromise allowing for simplifying the calculation and making results analytical and transparent for further interpretation. The Gaussian process is defined completely by its first and second central moments. The price for the simplification of the model and for the possibility to have an analytical solution is a non-zero probability to have the negative values for a covariate. Thus, the BBM model in the form of the Gaussian process does not accurately represent the system with a small amount of intermediate cells, and this is a limitation of the BBM model. In practice it does not create the problems due to the following reasons.

First, the covariates used in our model (also known as cancer hallmarks) represent the fractions of cells with broken (or maybe activated) cellular processes preventing (or maybe promoting) carcinogenesis. Occurrence of new mutations or adverse epigenetic events in the pool of susceptible cells and their forthcoming elimination is not a rare event. There processes can be in a dynamic equilibrium resulting in a unimodal distribution of cells with

detectable barrier failure. There exist both theoretical analyses [61] and experimental results [53] supporting it. Therefore, we can speculate about the norm of cells with a broken barrier mechanism in a population. Thus, we may expect the Gaussian distribution of covariate (or approximate them by the Gaussian distribution) defined in this way in a population. The norm is the mode of such a distribution. Note, that the models based on Gaussian approximation incorporated the notion of the norm are successfully used in biodemography and aging research [15–18,60,62]. Different modifications of SPM with age-dependent parameters (such as age-dependent norms) were extensively analyzed in simulation studies which showed identifiability and feasibility of the models [15–18,60,62] and were used in analyses of real data [62–67].

Second, the BBM model is designed to deal with the longitudinal data; therefore, it describes the covariate dynamics by a conditional Gaussian distribution, i.e., the Gaussian distribution conditional on previous covariate measurements. At the time of a measurement, the variance of a covariate is set to zero, and only during a long period without additional measurements could it result in a covariate distribution with non-vanishing fractions of covariates with negative values. The model for incidence rate only deals with first and second central moments which are integration characteristics, and possible situations in data analysis when the first moment is negative, may manifest problems in parameter estimates (e.g., limited statistical power or too large a time period between measurements).

4.4. Further model generalization

In this paper, we restricted ourselves to standard assumptions prevailing in currently adapted theories of carcinogenesis considering cell initiation, promotion, and conversion to be the major events in cancer development. The detailed specification of a model could result in over-parameterization (such as that the newly introduced parameters could be non-identifiable). Moreover, further detailing requires a new, more sophisticated, mathematical formalism, often associated with technical difficulties. The model developed in this paper is capable of overcoming these problems. First, our approach allows the use of traditional sets of epidemiological data together with series of biologically motivated measurements that help in covering a deficit in amount and diversity of available experimental information. Second, the suggested methodology is based on the well-developed theory of the diffusion type stochastic differential equations, which creates a good background for further extensions of this model. Among the prospective methodological generalizations of the model is the use of age-dependent model parameters representing aging/ontogeny-related processes in an organism that could modulate the tumor development at all stages (from the appearance of a transformed cell to a clinical manifestation of cancer): while the inefficiency of barrier mechanisms can be a plausible explanation for an initial cell transformation, the probability of survival of a malignant cell and/or latent tumors in constantly changing tissues of an aging body is likely to be influenced by additional factors, such as cell microenvironment, adaptation of tissues to a specific treatment/disease and, of course, background aging/ontogeny-related changes in the body per se.

Several specific directions for further generalizations of the model are concerned with an improvement of the predictive power of the model and analyses of possible uncertainties and biases. One important step is to develop a more individualized model of development of leukemia and solid cancers. Technically, this means that the risk of solid cancer or leukemia has to be predicted at a certain degree based on specific individual measurements and knowledge of population characteristics. An important component of such prognoses is the estimations of statistical uncertainties and systematic biases. The second step (e.g., necessary for the leukemia model) is to include the model of different stages of differentiation of blood cells, i.e., to describe the process from red-bone stem cells to cells of peripheral blood. This study is necessary because the conclusions about red bone stem cells made by measuring barrier states of peripheral leukocytes can be biased despite the same genetics and maybe epigenetics. The first step in constructing such a joint model was done by Akushevich et al. [52,68]. Another possible generalization step is to extend the set of the mechanisms considered as barriers, which also play an important role in preventing carcinogenesis, e.g., lack of telomerase production, cells' growth arrest, etc. Finally, while using and generalizing the concept of barrier mechanisms, it is possible to develop the model of IR-induced genomic instability, a promising mechanism of carcinogenesis [42,43,69,70]. The effects of genomic instability are naturally incorporated into the BBM model because the states of the barrier mechanisms are simultaneously the biomarkers of genomic instability. All these generalizations have to and will be supported by the corresponding methodological development.

Acknowledgment

This work was supported by NIA/NIH Grants R01AG028259, R01AG032319, R01AG030198

Appendix A

The hazard rate $\bar{h}(t)$ for the reduced set of parameters v_0, Δ, C, C_C has to be obtained as a solution of the system (9) and (10). Since the hazard rate is simply $\bar{h}(t) = m(t)^* \bar{\mu}_1$, where the 5-dimensional vector $m(t)$ is the solution of the vector differential equations $dm(t)/dt = a_0 + a_1 m(t) - \gamma(t)\bar{\mu}_1$ ($m(0) = \mu_0$), which in turn depends on the $\gamma(t)$ being the solution of the matrix system of equations $d\gamma(t)/dt = a_1\gamma^*(t) + \gamma(t)a_1^*(t) + bb^*(\gamma(0) = 0)$, we start with an analysis of the latter matrix system. An analytical solution of all these differential equations is possible, and the first step is the solution to 15 independent equations for symmetric matrix $\gamma(t)$. To simplify the matrix element notation, we use numerical indices for barrier states: 1-*BC*, 2-*AC*, 3-*AB*, 4-*B*, 5-*A*. Note, first, that 5 matrix elements $\gamma_{12}(t)$, $\gamma_{13}(t)$, $\gamma_{15}(t)$, $\gamma_{23}(t)$, and $\gamma_{24}(t)$ (and the elements symmetrical to them) are expressed in terms of themselves, and therefore they are equal to zero for whole time period because of initial condition $\gamma(0) = 0$. For example equation for $\gamma_{12}(t)$ is $\gamma'_{12}(t) = 2\Delta\gamma_{12}(t)$, and equation for $\gamma_{24}(t)$ is $\gamma'_{24}(t) = 2\Delta\gamma_{24}(t) + v_0 C\gamma_{12}(t) + v_0 C_c\gamma_{23}(t)$. The equations for the remaining 10 matrix elements are

$$\gamma'_{11}(t) = 2\Delta\gamma_{11}(t) + b_1,$$

$$\gamma'_{14}(t) = 2\Delta\gamma_{14}(t) + v_0 C\gamma_{11}(t),$$

$$\gamma'_{22}(t) = 2\Delta\gamma_{22}(t) + b_2,$$

$$\gamma'_{25}(t) = 2\Delta\gamma_{25}(t) + v_0 C\gamma_{22}(t),$$

$$\gamma'_{33}(t) = b_3,$$

$$\gamma'_{34}(t) = \Delta\gamma_{34}(t) + v_0 C_c\gamma_{33}(t),$$

$$\gamma'_{35}(t) = \Delta\gamma_{35}(t) + v_0 C_c\gamma_{33}(t),$$

$$\gamma'_{44}(t) = 2\Delta\gamma_{44}(t) + 2v_0 C\gamma_{14}(t) + 2v_0 C_c\gamma_{34}(t) + b_4,$$

$$\gamma'_{45}(t) = 2\Delta\gamma_{45}(t) + v_0 C_c(\gamma_{35}(t) + \gamma_{34}(t)),$$

$$\gamma'_{55}(t) = 2\Delta\gamma_{55}(t) + 2v_0 C\gamma_{25}(t) + 2v_0 C_c\gamma_{35}(t) + b_5.$$

These equations can be solved one by one. Evidently, the solutions accounting for the initial conditions $\gamma(0) = 0$ can be obtained in analytical form. They are

$$\gamma_{11}(t) = \frac{b_1}{2\Delta}(e^{2\Delta t} - 1),$$

$$\gamma_{14}(t) = \frac{v_0 C b_1}{4\Delta^2}((2\Delta t - 1)e^{2\Delta t} + 1),$$

$$\gamma_{22}(t) = \frac{b_2}{2\Delta}(e^{2\Delta t} - 1),$$

$$\gamma_{25}(t) = \frac{v_0 C b_2}{4\Delta^2}((2\Delta t - 1)e^{2\Delta t} + 1),$$

$$\gamma_{33}(t) = b_3 t,$$

$$\gamma_{34}(t) = \gamma_{35}(t) = \frac{v_0 C_c b_1}{\Delta^2}(e^{\Delta t} - \Delta t - 1),$$

$$\gamma_{44}(t) = \frac{b_4}{2\Delta}(e^{2\Delta t} - 1) + \frac{v_0^2 C^2 b_1}{4\Delta^3}((2\Delta^2 t^2 - 2\Delta t + 1)e^{2\Delta t} - 1) + \frac{v_0^2 C_c^2 b_3}{2\Delta^3}(e^{2\Delta t} - 4e^{\Delta t} + 2\Delta t + 3),$$

$$\gamma_{45}(t) = \frac{v_0^2 C_c^2 b_3}{2\Delta^3}(e^{2\Delta t} - 4e^{\Delta t} + 2\Delta t + 3),$$

$$\gamma_{55}(t) = \frac{b_5}{2\Delta}(e^{2\Delta t} - 1) + \frac{v_0^2 C^2 b_2}{4\Delta^3}((2\Delta^2 t^2 - 2\Delta t + 1)e^{2\Delta t} - 1) + \frac{v_0^2 C_c^2 b_3}{2\Delta^3}(e^{2\Delta t} - 4e^{\Delta t} + 2\Delta t + 3).$$

These results are used in the R.H.S of equations for $m(t)$, which can also be solved analytically. The results are presented using a simplified notation (13):

$$m_1(t) = Em_{10} + \frac{\phi}{\Delta}(E-1) - \frac{\varepsilon_1 b_1}{4\Delta}(2(E-1)^2 + \varepsilon(-(E-1)(3E-1) + 2\tau E^2)),$$

$$m_3(t) = m_{30} + t\phi + \frac{\varepsilon\varepsilon_2 b_3}{\Delta}(1 - 2E + (1+\tau)^2),$$

$$m_4(t) = Em_{40} + \varepsilon_1 \tau E m_{10} + \varepsilon_2 (E-1)m_{30} + \frac{\phi}{\Delta}((\varepsilon_2 - \varepsilon_1)(E-1) + \tau(\varepsilon_1 E - \varepsilon_2)) - \frac{\varepsilon_1^2 b_1}{4\Delta}(-(E-1)(E-3) + 2\tau E(E-2) + 2\varepsilon(1 - E(1-\tau))^2) - \frac{\varepsilon\varepsilon_2^2 b_3}{\Delta}(E - 1 - \tau)^2 - \frac{\varepsilon b_4}{2\Delta}(E-1)^2,$$

Solution $m_2(t)$ is obtained from $m_1(t)$ by substitution $b_1 \to b_2$ and $m_{10} \to m_{20}$ and $m_5(t)$ is obtained from $m_4(t)$ by substitution $b_1 \to b_2$, $b_4 \to b_5$, $m_{10} \to m_{20}$ and $m_{40} \to m_{50}$.

Finally, the obtained expressions for $m(t)$ are combined according to

$$\bar{h}(t) = \mu_{BA} m_1(t) + \mu_{AB} m_2(t) + \mu_{BAC} m_4(t) + \mu_{ABC} m_5(t),$$

to obtain the result for the conditional hazard in the form of (12).

References

[1] C.O. Nordling, A new theory on cancer-inducing mechanism, Br. J. Cancer 7 (1953) 68.
[2] P. Armitage, R. Doll, The age distribution of cancer and a multi-stage theory of carcinogenesis, Br. J. Cancer 8 (1954) 1.
[3] S. Moolgavkar, D. Krewski, M. Schwarz, Mechanisms of carcinogenesis and biologically based models for estimation and prediction of risk, in: S. Moolgavkar, D. Krewski, L. Zeise, E. Cardis and H. Møller (Eds.), Quantitative Estimation and Prediction of Human Cancer Risks, Scientific publications No. 131, International Agency for Research on Cancer, Lyon, 1999, pp. 179–237.
[4] I.M. van Leeuwen, C. Zonneveld, From exposure to effect: a comparison of modeling approaches to chemical carcinogenesis, Mutat. Res. 489 (1) (2001) 17;
I.M. van Leeuwen, C. Zonneveld, From exposure to effect: a comparison of modeling approaches to chemical carcinogenesis, Erratum in: Mutat. Res. 511 (1) (2002) 87.
[5] W.F. Heidenreich, E.G. Luebeck, W.D. Hazelton, H.G. Paretzke, S.H. Moolgavkar, Multistage models and the incidence of cancer in the cohort of atomic bomb survivors, Radiat. Res. 158 (5) (2002) 607.
[6] K.G. Arbeev, S.V. Ukraintseva, L.S. Arbeeva, A.I. Yashin, Mathematical models for human cancer incidence rates, Demograph. Res. 12 (10) (2005) 237.
[7] M. Little, W. Heidenreich, S. Moolgavkar, H. Schöllnberger, D. Thomas, Systems biological and mechanistic modelling of radiationinduced cancer, Radiat. Environ. Biophys. 47 (2007) 39.
[8] A.I. Yashin, K.G. Manton, J.W. Vaupel, Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables, Theor. Pop. Biol. 27 (1985) 154.
[9] A.I. Yashin, K.G. Manton, Effects of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies, Stat. Sci. 12 (1997) 20.
[10] W.Y. Tan, Probability distribution of the number of initiated cells of carcinogenesis under prevention, Math. Comput. Model. 41 (2005) 1403.
[11] G.A. Veremeyeva, I.V. Akushevich, S.V. Ukraintseva, A.I. Yashin, S.B. Epifanova, E.A. Blinova, A.V. Akleyev, A new approach to individual prognostication of cancer development under conditions of chronic radiation exposure, Int. J. Low Rad. 7 (2010) 53.
[12] D. Hanahan, R.A. Weinberg, The hallmarks of cancer, Cell 100 (2000) 57.
[13] Surveillance, Epidemiology, and End Results (SEER), Program, <http://www.seer.cancer.gov>, 2007.
[14] M.A. Woodbury, K.G. Manton, A random walk model of human mortality and aging, Theor. Pop. Biol. 11 (1977) 37.
[15] I. Akushevich, A. Kulminski, K. Manton, Life tables with covariates: dynamic model for nonlinear analysis of longitudinal data, Math. Pop. Stud. 12 (2005) 51.
[16] A.I. Yashin, K.G. Arbeev, I. Akushevich, A. Kulminski, L. Akushevich, S.V. Ukraintseva, Stochastic model for analysis of longitudinal data on aging and mortality, Math. Biosci. 208 (2) (2007) 538.
[17] A.I. Yashin, K.G. Arbeev, I. Akushevich, A. Kulminski, L. Akushevich, S.V. Ukraintseva, Model of hidden heterogeneity in longitudinal data, Theor. Pop. Biol. 73 (1) (2008) 1.
[18] K.G. Arbeev, I. Akushevich, A.M. Kulminski, L.S. Arbeeva, L. Akushevich, S.V. Ukraintseva, I.V. Culminskaya, A.I. Yashin, Genetic model for longitudinal studies of aging, health, and longevity and its potential application to incomplete data, J. Theor. Biol. 258 (2009) 103.
[19] A.I. Yashin, Dynamics in survival analysis: conditional Gaussian property vs. Cameron–Martin formula, in: N.V. Krylov, R.Sh. Lipster, A.A. Novikov (Eds.), Statistics and Control of stochastic processes, Springer, New York, 1985, p. 446.
[20] I. Akushevich, K.G. Manton, A. Kulminski, M. Kovtun, J. Kravchenko, A. Yashin, Population models for the health effects of ionizing radiation, Radiat. Biol. Radioecol. 46 (2006) 663.
[21] UNSCEAR 2000. United Nations scientific committee on the effects of atomic radiation, Health Phys. 80 (3) (2001) 291.
[22] W.F. Heidenreich, H.G. Paretzke, The two-stage clonal expansion model as an example of a biologically based model of radiation-induced cancer, Radiat. Res. 156 (5 Pt 2) (2001) 678.
[23] S.H. Moolgavkar, A. Dewanji, D.J. Venzon, A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor, Risk Anal. 8 (3) (1988) 383.
[24] S.H. Moolgavkar, E.G. Luebeck, Two-event model for carcinogenesis: biological, mathematical and statistical considerations, Risk Anal. 10 (1990) 323.
[25] W.E. Heidenreich, P. Jacob, H.G. Paretzke, Exact solutions of the clonal expansion model and their application to the incidence of solid tumors of atomic bomb survivors, Radiat. Environ. Biophys. 36 (1997) 45.
[26] B. Øksendal, Stochastic Differential Equations: An Introduction with Applications, sixth ed., Springer, Berlin, 2003.
[27] M.P. Alison (Ed.), The Cancer Handbook, 2nd ed., vol. 2, Springer, 2007.
[28] A.V. Akleyev, G.A. Veremeyeva, A.V. Vozilova, Remote effects at cell and subcell level in the hemopoietic system after chronic radiation exposure in man, Radiat. Biol. Radioecl. 46 (2006) 519.
[29] R. Meza, W.D. Hazelton, G.A. Colditz, S.H. Moolgavkar, Analysis of lung cancer incidence in the nurses' health and the health professionals' follow-up studies using a multistage carcinogenesis model, Cancer Causes Control 19 (2008) 317.
[30] J. Kravchenko, I. Akushevich, V.L. Seewaldt, A.P. Abernethy, H.K. Lyerly, Breast cancer as heterogeneous disease: contributing factors and carcinogenesis mechanisms, Breast Cancer Res. Treat. 128 (2011) 483.
[31] C. Colijn, M.C. Mackey, A mathematical model of hematopoiesis. I. Periodic chronic myelogenous leukemia, J. Theor. Biol. 237 (2005) 117.
[32] S.H. Moolgavkar, D.J. Venzon, Two-event models for carcinogenesis: incidence curves for childhood and adult tumors, Math. Biosci. 47 (1979) 55.

[33] S.H. Moolgavkar, E.G. Luebeck, M. de Gunst, R.E. Port, M. Schwarz, Quantitative analysis of enzyme-altered foci in rat hepatocarcinogenesis experiments. I. Single agent regimen, Carcinogenesis 11 (8) (1990) 1271.
[34] S.H. Moolgavkar, F.T. Cross, E.G. Luebeck, G.E. Dagle, A two-mutation model for radon-induced lung tumors in rats, Rad. Res. 121 (1990) 28.
[35] A. Kopp-Schneider, C.J. Portier, Birth and death differentiation rates of papillomas in mouse skin, Carcinogenesis 13 (1992) 973.
[36] E.G. Luebeck, S.B. Curtis, F.T. Cross, S.H. Moolgavkar, Two-stage model of radon-induced malignant lung tumors in rats: effects of cell killing, Rad. Res. 145 (1996) 163.
[37] S.H. Moolgavkar, A. Knudson, Mutation and cancer: a model for human carcinogenesis, Natl. Cancer Inst. 66 (1981) 1037.
[38] S.H. Moolgavkar, Model for human carcinogenesis: action of environmental agents, Environ. Health Perspect. 50 (1983) 285.
[39] S.H. Moolgavkar, E.G. Luebeck, D. Krewski, J.M. Zielinski, Radon, cigarette smoke, and lung cancer: a reanalysis of the colorado plateau uranium miners' data, Epidemiology 4 (3) (1993) 204.
[40] S.H. Moolgavkar, E.G. Luebeck, Multistage carcinogenesis: population-based model for colon cancer, J. Nutl. Cancer Inst. 84 (1992) 610.
[41] M.P. Little, Are two mutations sufficient to cause cancer? some generalization to the two-mutation model of carcinogenesis of moolgavkar, venzon and knudson, and of the multistage model of armitage and doll, Biometrics 51 (1995) 1278.
[42] M.P. Little, P. Vineis, G. Li, A stochastic carcinogenesis model incorporating multiple types of genomic instability fitted to colon cancer data, J. Theor. Biol. 254 (2008) 229.
[43] M.P. Little, E.G. Wright, A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data, Math. Biosci. 183 (2003) 111.
[44] E.G. Luebeck, S.H. Moolgavkar, Multistage carcinogenesis and the incidence of colorectal cancer, PNAS 99 (2002) 15095.
[45] R. Meza, J. Jeon, S.H. Moolgavkar, E.G. Luebeck, Age-specific incidence of cancer: phases, transitions, and biological implications, PNAS 105 (2008) 16284.
[46] S.H. Moolgavkar, R. Meza, J. Turim, Pleural and peritoneal mesotheliomas in SEER: age effects and temporal trends 1973–2005, Cancer Causes Control 20 (2009) 935–944.
[47] M.A. Smith, T. Chen, R. Simon, Age-specific incidence of acute lymphoblastic leukemia in US children: in utero initiation model, JNCI 89 (1997) 1542.
[48] W.Y. Tan, X.W. Yan, A new stochastic and state space model of human colon cancer incorporating multiple pathways, Biol. Direct 5 (2010) 26.
[49] W.D. Hazelton, M.S. Clements, S.H. Moolgavkar, Multistage carcinogenesis and lung cancer mortality in three cohorts., Cancer Epidemiol. Biomark. Prevention 14 (2005) 1171.
[50] R. Bhattacharya, E. Waymire, Stochastic Processes with Applications, Wiley, New York, 1990.
[51] C.W. Gardiner, Stochastic Methods: For the Natural and Social Sciences, 4th ed., Springer, 2009.
[52] I.V. Akushevich, G.A. Veremeyeva, G.P. Dimov, S.V. Ukraintseva, K.G. Arbeev, A.V. Akleyev, A.I. Yashin, Modeling deterministic effects in hematopoietic system caused by chronic exposure to ionizing radiation in large human cohorts, Health Phys. J. 99 (2010) 322.
[53] G. Veremeyeva, I. Akushevich, T. Pochukhailova, E. Blinova, T. Varfolomeyeva, O. Ploshchanskaya, O. Khudyakova, A. Vozilova, O. Kozionova, A. Akleyev, Long-term cellular effects in humans chronically exposed to ionizing radiation, Health Phys. J. 99 (2010) 337.
[54] B. Vogelstein, K.W. Kinzler, Cancer genes and the pathways they control, Nat. Med. 10 (2004) 789.
[55] E.L. Wynder, J.E. Muscat, The changing epidemiology of smoking and lung cancer histology, Environ. Health Perspect. 103 (Suppl) (1995) 143.
[56] L.D. Wood, D.W. Parsons, S. Jones, J. Lin, T. Sjoblom, T. Barber, et al., The genomic landscapes of human breast and colorectal cancers, Science 318 (2007) 1108.
[57] T. van Dyke, T. Jacks, Cancer modeling in the modern era: progress and challenges, Cell 108 (2002) 135.
[58] M. Kovtun, I. Akushevich, K.G. Manton, H.D. Tolley, Linear latent structure analysis: mixture distribution models with linear constraints, Stat. Methodol. 4 (2007) 90.
[59] I. Akushevich, M. Kovtun, K.G. Manton, A.I. Yashin, Linear latent structure analysis and modelling of multiple categorical variables, Comput. Math. Methods Med. 10 (3) (2009) 203.
[60] A.I. Yashin, I. Akushevich, K. Arbeev, A. Kulminski, S. Ukraintseva, Joint analysis of health histories, physiological state, and survival, Math. Pop. Stud. 18 (2011) 207.
[61] A. Dewanji, E.G. Luebeck, S.H. Moolgavkar, A generalized Luria–Delbrück model, Math. Biosci. 197 (2) (2005) 140.
[62] K.G. Arbeev, S.V. Ukraintseva, I. Akushevich, A.M. Kulminski, L.S. Arbeeva, L. Akushevich, I.V. Culminskaya, A.I. Yashin, Age trajectories of physiological indices in relation to healthy life course, Mech. Ageing Develop. 132 (3) (2011) 93.
[63] A.I. Yashin, K.G. Arbeev, A. Kulminski, I. Akushevich, L. Akushevich, S.V. Ukraintseva, Health decline, aging and mortality: how are they related, Biogerontology 8 (3) (2007) 291.
[64] A.I. Yashin, K.G. Arbeev, A. Kulminski, I. Akushevich, L. Akushevich, S.V. Ukraintseva, What age trajectories of cumulative deficits and medical costs tell us about individual aging and mortality risk: findings from the NLTCS-Medicare data, Mech. Aging Develop. 129 (4) (2008) 191.
[65] A.I. Yashin, S.V. Ukraintseva, K.G. Arbeev, I. Akushevich, L.S. Arbeeva, A.M. Kulminski, Maintaining physiological state for exceptional survival: what is the normal level of blood glucose and does it change with age, Mech. Ageing Develop. 130 (9) (2009) 611.
[66] A.I. Yashin, K.G. Arbeev, I. Akushevich, S.V. Ukraintseva, A. Kulminski, L.S. Arbeeva, I. Culminskaya, Exceptional survivors have lower age trajectories of blood glucose: lessons from longitudinal data, Biogerontology 11 (3) (2010) 257.
[67] A.I. Yashin, K.G. Arbeev, S.V. Ukraintseva, I. Akushevich, A. Kulminski, Patterns of aging related changes on the way to 100: an approach to studying aging, mortality, and longevity from longitudinal data, in: 2011 Living to 100 Monograph, Society of Actuaries Monograph M-Ll11-1, Schaumburg, IL, 2011, <http://www.soa.org/library/monographs/life/living-to-100/2011/2011-toc.aspx>.
[68] I. Akushevich, G.A. Veremeyeva, G.P. Dimov, S.V. Ukraintseva, K.G. Arbeev, A.V. Akleyev, A.I. Yashin, Modeling hematopoietic system response caused by chronic exposure to ionizing radiation, Radiat. Environ. Biophys. 50 (2011) 299.
[69] M.A. Nowak, N.L. Komarova, A. Sengupta, P.V. Jallepalli, IeM. Shih, B. Vogelstein, C. Lengauer, The role of chromosomal instability in tumor initiation, Proc. Natl. Acad. Sci. USA 99 (2002) 16226.
[70] M.P. Little, G. Li, Stochastic modelling of colon cancer: is there a role for genomic instability, Carcinogenesis 28 (2007) 479.