Mathematical Biosciences and Engineering

MBE, 21(1): 1186–1202.
DOI: 10.3934/mbe.2024050
Received: 15 October 2023
Revised: 27 November 2023
Accepted: 30 November 2023
Published: 25 December 2023

http://www.aimspress.com/journal/MBE

Research article

Mathematical modeling the gene mechanism of colorectal cancer and the effect of radiation exposure

Lingling Li[1,2,*], Yulu Hu[1], Xin Li[1] and Tianhai Tian[3]

[1] School of Science, Xi'an Polytechnic University, Xi'an 710048, China
[2] School of Mathematics and Statistics, Shaanxi Normal University, Xi'an 710048, China
[3] School of Mathematics, Monash University, Melbourne Vic 3800, Australia

* **Correspondence:** Email: linglinglimath@163.com.

Abstract: Cancer is the result of continuous accumulation of gene mutations in normal cells. The number of mutations is different in different types of cancer and even in different patients with the same type of cancer. Therefore, studying all possible numbers of gene mutations in malignant cells is of great value for the understanding of tumorigenesis and the treatment of cancer. To this end, we applied a stochastic mathematical model considering the clonal expansion of any premalignant cells with different mutations to analyze the number of gene mutations in colorectal cancer. The age-specific colorectal cancer incidence rates from the Surveillance, Epidemiology and End Results (SEER) registry in the United States and the Life Span Study (LSS) in Nagasaki and Hiroshima, Japan are chosen to test the reasonableness of the model. Our fitting results indicate that the transformation from normal cells to malignant cells may undergo two to five driver mutations for colorectal cancer patients without radiation-exposed environment, two to four driver mutations for colorectal cancer patients with low level radiation-exposure, and two to three driver mutations for colorectal cancer patients with high level radiation-exposure. Furthermore, the net growth rate of the mutated cells with radiation-exposure was is higher than that of the mutated cells without radiation-exposure for the models with two to five driver mutations. These results suggest that radiation environment may affect the clonal expansion of cells and significantly affect the development of tumors.

Keywords: mathematical modeling; colorectal cancer; incidence rate; driver mutation; radiation exposure

1. Introduction

Colorectal cancer is one of the common malignant tumors that seriously endangers human health. Statistical data showed that colorectal cancer is the second largest cancer in the world and the second death rate in developed countries [1,2]. In recent years, the incidence rate of colorectal cancer has gradually increased in young and adult groups, which is closely related to the changes in people's lifestyles [3–5]. However, age-specific incidence rate of colorectal cancer showed a significant decrease in the age group above 85 years [6]. In addition, Jones et al. found that it took more than ten years for benign tumors to develop into advanced cancer in the colorectal and two years to obtain the metastatic ability [7]. However, most cancer patients are already in the advanced stage at the time of diagnosis. Therefore, the earlier the disease is found, the higher survival rate patients with colorectal cancer have [8]. Studying the mechanism of gene mutation is vital for the early diagnosis in the development of cancer.

Cancer is essentially a gene-related disease, but a single gene mutation cannot cause tumor [9]. Researches suggested that there is a large number of gene mutations in the process of tumorigeneses [9–11]. Among them, however, only a small fraction of mutations are driver mutations that cause the selective growth advantage to cells, leading to the development of tumors [9–12]. Therefore, many mathematical models considering gene mutation were developed to analyze the risk of cancer. The earliest cancer model could be traced back to the work of Armitage and Doll [13]. They proposed the multistage cancer model and fitted the age-specific mortality data of colorectal cancer, gastric cancer and other cancers by using their established model. Later, there was evidence to show that the clonal expansion of cells played an important role in the development of cancer. For this reason, researchers developed the models with clonal expansion of mutant cells, which could better explain the risk of colorectal cancer than the model without clonal expansion of cells [14–17]. In addition, Lang et al. studied the cancer model based on two branching processes and obtained the growth and metastasis rate of adenoma by fitting the data of colorectal cancer incidence rate and adenoma size [18]. However, these studies did not consider the specific gene information. To address this issue, Paterson et al. proposed a five-step branching process model, including suppressor genes APC, TP53 and oncogene KRAS, to explore the detailed sequence of these gene mutations in the colorectal cancer [19,20]. Nevertheless, only 15% colorectal cancer patients have mutations in suppressor genes APC, TP53 and oncogene KRAS [21,22]. Thus, their result was not applicable to all colorectal cancer patients.

In our work, we develop an any-step branching process model with clonal expansion to study the number of driver gene mutations in colorectal cancer instead of using a branching process model with fixed step, which allows us to study the influence of tumor heterogeneity. The model with any step involving clonal expansion of cells matches the age-specific colorectal cancer incidence rate from the Surveillance, Epidemiology and End Results registry (SEER) in the United States data and the Life Span Study (LSS) in Nagasaki and Hiroshima, Japan with different dose level exposure, which obtains all possible numbers of driver gene mutations that can be used to explain the progression of colorectal cancer. Moreover, we analyze the estimated parameters of the model to verify the reasonableness of the fitting results. Our study results can be used to identify the influence of radiation exposure dose on the risk of colorectal cancer.

2. Materials and methods

2.1. The data

The SEER and LSS databases are two main data sources for studying cancer, which can be obtained from https://seer.cancer.gov/ and https://www.rerf.or.jp/en/, respectively. Here, we choose colorectal cancer incidence rate data from the SEER registry during the years 1973–2013 and from the LSS dataset during the years 1958 to 1998, respectively. For the LSS dataset, we divide into two categories with colon dose level. The data with less than 0.1 Gy is considered as low dose level exposure, and other data is considered as high dose level exposure [23]. This data is classified by the 5-year age groups (namely, age 0–4, 5–9, …, 75–80). In particular, we find that the incidence rate of colorectal cancer is almost zero before the age of 25 years in the LSS data. For the LSS data, we fit the data from 25–80 years of age. In our analyses, the numbers of patients over 80 years of age are ignored since they decline rapidly. In addition, we assume that the latent period of colorectal cancer is 5 years, which means that the tumor will be detected clinically after 5 years when a persistent malignant tumor cell appears. Table 1 gives the cases and total population in each age group from the SEER and the LSS with low (<0.1 Gy) and high dose (≥0.1 Gy) exposure data.

Table 1. Age-specific colorectal cancer incidence data from the SEER registry during 1973–2013 and the LSS dataset during 1958–1998.

	SEER		LSS			
Age	Cases	Person years	<0.1 Gy		≥0.1 Gy	
			Cases	Person years	Cases	Person years
0–4	3	72,394,578	0	0	0	0
5–9	3	72,210,122	0	0	0	0
10–14	55	74,259,274	0	4012	0	792
15–19	188	75,390,703	0	32,374	0	6952
20–24	589	76,696,109	0	51,398	0	10,865
25–29	1376	79,981,477	1	73,971	2	15,072
30–34	2893	79,416,642	3	100,004	1	20,405
35–39	5396	74,888,049	11	117,356	1	24,149
40–44	10,084	70,349,908	14	134,452	4	27,601
45–49	17,957	64,617,689	28	153,632	9	31,443
50–54	31,526	58,921,791	76	173,091	29	35,664
55–59	42,381	51,312,063	130	171,197	39	35,682
60–64	54,754	43,156,922	195	163,996	41	34,579
65–69	67,393	35,513,414	243	148,859	65	31,941
70–74	74,473	28,508,602	219	118,754	59	25,794
75–79	74,301	22,311,393	193	88,718	51	19,102

2.2. Mathematical model

It is very valuable to explore all possible numbers of driver genes in the colorectal cancer for the diagnose and therapy of a tumor. Here, we describe the any-step branching process model with clonal expansion of cells, which assumes that normal cells undergo several rate-limiting events before developing into malignant cells. Let N denote normal cells, P_i mutated cells with i mutation(s) where $i=1,2,\cdots,k-1$, and P_k malignant cells. The normal cells (N) and mutated cells (P_i) undergo mutations at rates μ_N and μ_i, respectively. Furthermore, the mutated cells (P_i) undergo growth and death (or differentiation) at rates a_i and b_i, respectively, then the model with multiple branching processes is described as follows,

$$\begin{aligned}
&N\overset{\mu_N}{\rightarrow}P_1;\\
&P_i\overset{\mu_i}{\rightarrow}P_{i+1},\ i=1,2,\cdots,k-1;\\
&P_i\overset{a_i}{\rightarrow}P_i+P_i;\\
&P_i\overset{b_i}{\rightarrow}\varnothing.
\end{aligned}$$

In our model, we assume that the number of normal cells is constant and the probability of tumor is one once one malignant cell appears.

We consider the process from normal cells to a malignant cell and let $N(t)$, $P_i(t)$ $(i=1,2,\cdots,k-1)$ and $P_k(t)$ represent the number of normal cells, mutant cells with i mutation(s) and malignant tumor cells at time t, respectively. The following probability generating functions are defined for $\tau\le t$.

$$\begin{aligned}\Psi_0(p_1,p_2,\cdots,p_k;\tau,\mathrm{t})=\sum_{m_1,m_2,\cdots,m_k} prob\{&P_1(t)=m_1,P_2(t)=m_2,\cdots,P_{k-1}(t)=m_{k-1},P_k(t)=m_k\\ &|P_1(\tau)=0,P_2(\tau)=0,\cdots,P_{k-1}(\tau)=0,P_k(\tau)=0\}p_1^{m_1}p_2^{m_2}\cdots p_k^{m_k},\end{aligned}\tag{2.1}$$

$$\begin{aligned}\Psi_{\mathrm{i}}(p_i,p_{i+1},\cdots,p_k;\tau,\mathrm{t})=\sum_{m_i,m_{i+1},\cdots,m_k} prob\{&P_i(t)=m_i,P_{i+1}(t)=m_{i+1},\cdots,P_{k-1}(t)=m_{k-1},P_k(t)=m_k\\ &|P_i(\tau)=1,P_{i+1}(\tau)=0,\cdots,P_{k-1}(\tau)=0,P_k(\tau)=0\}p_i^{m_i}p_{i+1}^{m_{i+1}}\cdots p_k^{m_k},\end{aligned}\tag{2.2}$$

where $\Psi_0(1,1\cdots1;0,t)=1,\ \Psi_i(1,\cdots,1;0,t)=1$.

Using the probability generating Eqs (2.1) and (2.2), we can derive the probability of tumor as $P(t)=1-\Psi_0(1,\cdots1,0;0,t)$, and the expected number of mutated cells with i mutations is given by $E[P_i(t)]=\frac{\partial\Psi_i(p_1,\cdots,p_k;0,t)}{\partial p_i}\Big|_{(1,\cdots1)}$.

By the Kolmogorov forward equation, Equation (2.1) can be used to derive the following equation [24],

$$\begin{aligned}\frac{d\Psi_0}{dt}(p_1,p_2,\cdots,p_k;\tau,t)&=(p_1-1)N\mu_N\Psi_0(p_1,p_2,\cdots,p_k;\tau,t)\\ &+\sum_{i=1}^{k-1}[\mu_ip_ip_{i+1}+a_ip_i^2+b_i-(a_i+b_i+\mu_i)p_i]\frac{d\Psi_0(p_i,p_{i+1},\cdots,p_k;\tau,t)}{dp_i}.\end{aligned}\tag{2.3}$$

Taking the derivative of Eq (2.3) with respect to p_i and letting $(p_1,p_2,\cdots,p_k)=(1,1,\cdots,1)$, we have that

$$\begin{cases} \dfrac{dE[P_1(t)]}{dt} = \gamma_1 E[P_1(t)] + \mu_N N \\ \dfrac{dE[P_i(t)]}{dt} = \gamma_i E[P_i(t)] + \mu_{i-1} E[P_{i-1}(t)] \quad 2 \le i \le k-1 \end{cases} . \tag{2.4}$$

Additionally, by the Kolmogorov backward equation, Equations (2.1) and (2.2) can be used to derive the following Equations [25,26],

$$\frac{d\Psi_0}{d\tau}(1,1,\cdots,1,0;\tau,t) = -\mu_N N \Psi_0(1,1,\cdots,1,0;\tau,t)[\Psi_1(1,1,\cdots,1,0;\tau,t) - 1], \tag{2.5}$$

for $i \le k-2$,

$$\begin{aligned} \frac{d\Psi_i}{d\tau}(1,\cdots,1,0;\tau,t) = (a_i + b_i + \mu_i)\Psi_i(1,\cdots,1,0;\tau,t) - a_i \Psi^2{}_i(1,\cdots,1,0;\tau,t) \\ -\mu_i \Psi_i(1,\cdots,1,0;\tau,t)\Psi_{i+1}(1,\cdots,1,0;\tau,t) - b_i, \end{aligned} , \tag{2.6}$$

and for $i = k-1$,

$$\frac{d\Psi_{k-1}}{d\tau}(1,0;\tau,t) = (a_{k-1} + b_{k-1} + \mu_{k-1})\Psi_{k-1}(1,0;\tau,t) - a_{k-1}\Psi^2{}_{k-1}(1,0;\tau,t) - b_{k-1}. \tag{2.7}$$

We take the derivative of Eqs (2.5)–(2.7) with respect to time t,

$$\begin{cases} \dfrac{d\Psi'_0}{d\tau}(1,1,\cdots,1,0;\tau,t) = -\mu_N N \{\Psi'_0(1,1,\cdots,1,0;\tau,t)[\Psi_1(1,1,\cdots,1,0;\tau,t) - 1] \\ \qquad + \Psi_0(1,1,\cdots,1,0;\tau,t)\Psi'_1(1,1,\cdots,1,0;\tau,t)\} \\ \dfrac{d\Psi'_i}{d\tau}(1,\cdots,1,0;\tau,t) = (a_i + b_i + \mu_i)\Psi'_i(1,\cdots,1,0;\tau,t) \\ \qquad - \Psi'_i(1,\cdots,1,0;\tau,t)[2a_i\Psi_i(1,\cdots,1,0;\tau,t) + \mu_i \Psi_{i+1}(1,\cdots,1,0;\tau,t)] \\ \qquad - \mu_i \Psi_i(1,\cdots,1,0;\tau,t)\Psi'_{i+1}(1,\cdots,1,0;\tau,t), \qquad i \le k-2 \\ \dfrac{d\Psi'_{k-1}}{d\tau}(1,0;\tau,t) = (a_{k-1} + b_{k-1} + \mu_{k-1})\Psi'_{k-1}(1,0;\tau,t) - 2a_{k-1}\Psi_{k-1}(1,0;\tau,t)\Psi'_{k-1}(1,0;\tau,t) \end{cases} \tag{2.8}$$

with the following boundary value condition

$$\begin{cases} \Psi_0(t,t) = 1 \\ \Psi'_0(t,t) = 0 \\ \Psi_i(t,t) = 1 \\ \Psi'_j(t,t) = 0, \quad 1 \le j \le k-2 \\ \Psi'_{k-1}(t,t) = -\mu_{k-1} \end{cases} , \tag{2.9}$$

where prime ‘′’ represents the derivation with respect to time t.

In order to convert the boundary value condition to the initial value condition [27], we let $s=t-\tau$, then $A(s,t)=\Psi_0(1,1,\cdots 1,0;\tau,t)$, $B(s,t)=\Psi_0'(1,1,\cdots,1,0;\tau,t)$, $C_i(s,t)=\Psi_i(1,\cdots,1,0;\tau,t)$, $D_i(s,t)=\Psi_i'(1,\cdots,1,0;\tau,t)$, $1\le i\le k-1$. Equations (2.5)–(2.9) can then be transformed into the following equations,

$$\begin{cases} \frac{dA}{ds}(s,t)=\mu_N NA(s,t)(C_1(s,t)-1) \\ \frac{dB}{ds}(s,t)=\mu_N N[B(s,t)(C_1(s,t)-1)+A(s,t)(D_1(s,t)] \\ \frac{dC_i}{ds}(s,t)=-(a_i+b_i+\mu_i)C_i(s,t)+a_iC^2{}_i(s,t)+\mu_iC_i(s,t)C_{i+1}(s,t)+b_i \\ \frac{dD_i}{ds}(s,t)=-(a_i+b_i+\mu_i)D_i(s,t)+D_i(s,t)[2a_iC_i(s,t)+\mu_iC_{i+1}(s,t)] \\ \qquad\qquad +\mu_iC_i(s,t)D_{i+1}(s,t), \quad i\le k-2 \\ \frac{dC_{k-1}}{ds}(s,t)=-(a_{k-1}+b_{k-1}+\mu_{k-1})C_{k-1}(s,t)+a_{k-1}C^2{}_{k-1}(s,t)+b_{k-1} \\ \frac{dD_{k-1}}{ds}(s,t)=-(a_{k-1}+b_{k-1}+\mu_{k-1})D_{k-1}(s,t)+2a_{k-1}C_{k-1}(s,t)D_{k-1}(s,t) \end{cases} \tag{2.10}$$

with the initial value condition

$$\begin{cases} A(0,t)=1 \\ B(0,t)=0 \\ C_i(0,t)=1 \\ D_j(0,t)=0, \quad 1\le j\le k-2 \\ D_{k-1}(0,t)=-\mu_{k-1} \end{cases} . \tag{2.11}$$

By the definitions of probability generating function (2.1), the probability and the risk function of malignant tumor cells appearing at time t are $P(t)=1-A(t,t)$ and $h(t)=-\frac{B(t,t)}{A(t,t)}$, respectively.

2.3. Parameter estimation

Given a set of observed cases $\{O_{\mathrm{t}}\}$, and the corresponding person-years $\{n_t\}$, we assume that colorectal cancer cases follow Poisson distribution, $\delta_t=n_th(t)$, where $h(t)$ is the incidence of colorectal cancer at time t, which depends on the parameter set $\Theta=(N\mu_N,a_i,b_i,\mu_i)$. Assuming that the observations of cases are independent, the likelihood function for observed cases set $\{O_t\}$ can be written as follows,

$$L(\Theta)=\prod_t \frac{e^{-\delta_t}\delta_t^{O_t}}{O_t!}, \tag{2.12}$$

then the negative log likelihood function yields

$$NLL(\Theta)=-\sum_t(-\delta_t+O_t\ln(\delta_t)-\ln(O_t!)). \tag{2.13}$$

The optimal values of model parameters are estimated by minimizing $NLL(\Theta)$. We choose the optimization routine *fminsearch* in MATLAB to determine the optimal values of model parameters by minimizing $NLL(\Theta)$. AIC is used to measure the goodness of fit of the model, $AIC=Deviance+2n$, where n denotes the number of model parameters.

In simulations, the mutation rates of cells are limited to less than 0.01 and the net growth rates of cells are set between the range of 0 to 0.5 [17,28]. The reason for this is that the premalignant tumor cells do not have the ability to multiply indefinitely and the probability of one gene mutation is $\mu_i t \ll 1$. In addition, we find that the total number of cells will be extremely large, which is impossible for premalignant cells. It is shown that the number of cells in the tissue will exceed 10^9 (the minimum clinically detectable value) at more than 20 years if the net multiplication rate is one per year, which is clearly unrealistic in biology [29], and the number of premalignant cells is quite small in the order of 10^3 in an adenoma of linear size 1 cm [30,31]. Furthermore, the net proliferation rate of cells with i mutations should not be greater than that of cells with $i+1$ mutations due to driver mutation conferring the selective growth advantage to cells; that is, $a_i-b_i<a_{i+1}-b_{i+1}$. Thus, we limit all net multiplication rates of premalignant to be less than 0.5.

3. Results

Tumor heterogeneity is an issue that cannot be ignored [32]. Therefore, the model with fixed stage is unsuitable for explaining the development of tumors. Here, we consider the model with any stage to explore the number of driver mutations in colorectal cancer. The incidence rate data of colorectal cancer at age-specific from SEER and LSS databases involving radiation exposure levels are used as the test system, and the models with different numbers of driver mutation are used to match this data. We give all fitting results of the model with any mutation number from two to eight, and other models with more than eight mutations do not have better fitting than those with less than eight mutations, which is in line with the viewpoint of reference [33]. The fitting results are displayed in Figures 1–3, which includes the relative error between real data and simulated data. The relative error is defined by $\text{relative error}=\frac{|\text{real data-simulated data}|}{\text{real data}}$. Combining the fitting results of the model and the corresponding errors between real data and simulated data, we find that the models with three to four driver mutations can better fit the SEER data than other models, especially for young patients. In addition, the model with two mutations is not appropriate for fitting the SEER data at ages of less than ten years by Figure 1(b). Table 2 gives the deviance and AIC of the fittings, which suggests that the model with three driver mutations is the optimal model to explain the colorectal cancer of SEER data, and two driver mutations is the best for explaining the development of colorectal cancer of LSS data for both low and high dose levels.

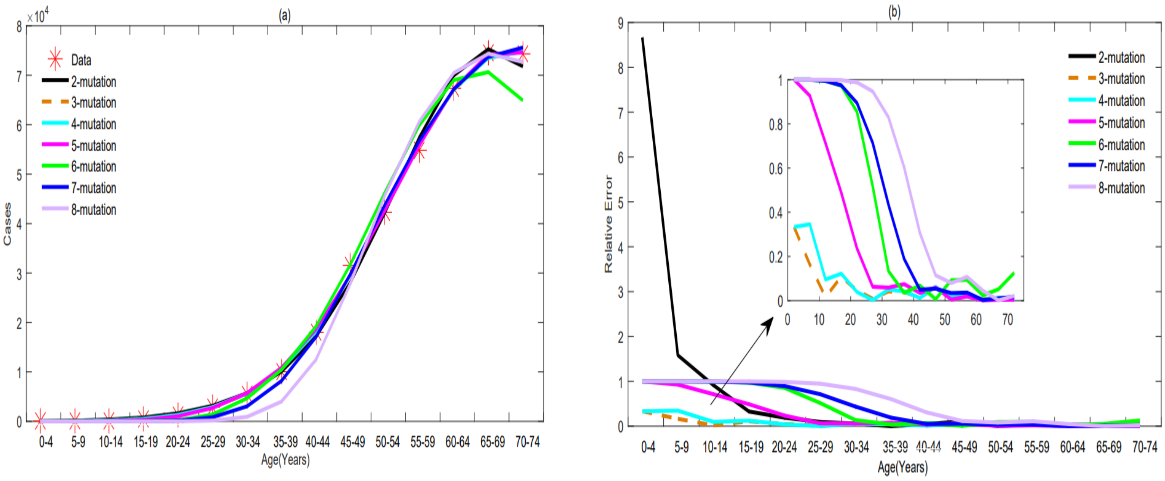

Figure 1. The fittings of the models with different numbers of mutations to age-specific colorectal cancer incidence rate from the SEER registry during 1973–2013 and the corresponding relative error between real data and simulated data.

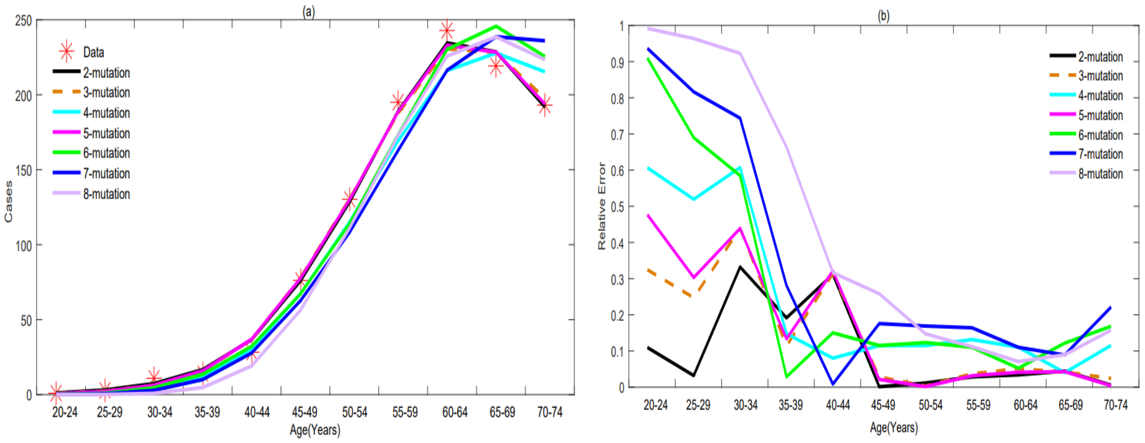

Figure 2. The fittings of the models with different numbers of mutations to age-specific colorectal cancer incidence data from the LSS with low-radiation exposure (colon dose <0.1 Gy) during 1958–1998 and the corresponding relative error between real data and simulated data.

Table 2. The deviance and AIC of the models with the number of mutations from two to eight for SEER data and LSS data with low radiation exposure and high radiation exposure.

k-stage			2	3	4	5	6	7	8
SEER		*Deviance*	1094.8	225.7	249.9	770.5	12,426	14,627	51,375
		AIC	1100.8	235.7	263.9	788.5	12,448	14,653	51,405
LSS	Low dose	*Deviance*	4.9	5.3	16.1	7.7	27.5	75.8	52.0
		AIC	10.9	15.3	30.1	25.7	49.5	101.8	82.0
	High-dose	*Deviance*	11.1	9.9	10.8	11.7	19.6	47.4	25.5
		AIC	17.1	19.9	24.8	29.7	41.6	73.4	55.5

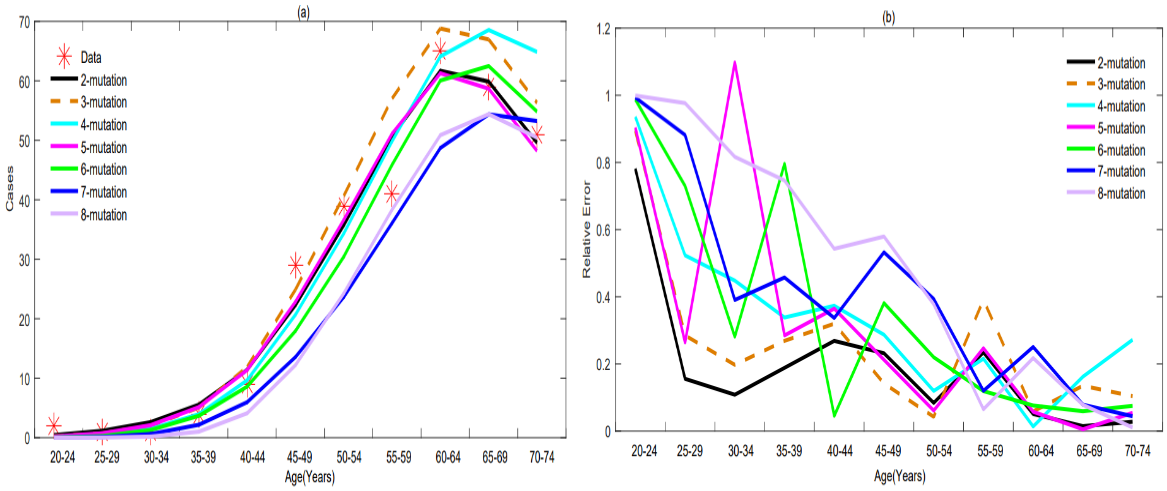

Figure 3. The fittings of the models with different numbers of mutations to age-specific colorectal cancer incidence data from the LSS with high-radiation exposure (colon dose ≥ 0.1 Gy) during 1958–1998 and the corresponding relative error between real data and simulated data.

Table 3. The net growth rates and mutation rates in the models with two to eight mutations for the SEER data without radiation exposure ($\gamma_i = a_i - b_i$).

	2	3	4	5	6	7	8
γ_1	0.129	0.030	0.017	0.032	0.119	0.052	0.033
γ_2		0.137	0.134	0.101	0.482	0.128	0.236
γ_3			0.134	0.102	0.495	0.300	0.255
γ_4				0.104	0.496	0.300	0.255
γ_5					0.499	0.461	0.387
γ_6						0.481	0.496
γ_7							0.500
$N\mu_N$	0.39	4.43	13.8	2.54	0.29	2.14	2.04
μ_1	4.19×10^{-7}	1.55×10^{-3}	6.47×10^{-4}	4.74×10^{-3}	3.91×10^{-3}	1.52×10^{-3}	1.46×10^{-3}
μ_2		1.97×10^{-6}	6.76×10^{-5}	1.53×10^{-3}	9.37×10^{-4}	5.81×10^{-3}	9.99×10^{-3}
μ_3			1.00×10^{-2}	9.62×10^{-3}	1.00×10^{-2}	1.00×10^{-2}	9.99×10^{-3}
μ_4				3.24×10^{-3}	6.91×10^{-3}	1.00×10^{-2}	9.97×10^{-3}
μ_5					1.00×10^{-2}	9.99×10^{-3}	1.00×10^{-2}
μ_6						9.16×10^{-3}	9.99×10^{-3}
μ_7							1.00×10^{-2}

Table 4. The net growth rates and mutation rates in the models with two to eight mutations for the LSS data with a low-radiation exposure level ($\gamma_i = a_i - b_i$).

	2	3	4	5	6	7	8
γ_1	0.142	0.049	0.019	0.121	0.115	0.048	0.025
γ_2		0.163	0.187	0.121	0.446	0.118	0.218
γ_3			0.188	0.347	0.495	0.300	0.255
γ_4				0.347	0.496	0.300	0.255
γ_5					0.500	0.335	0.387
γ_6						0.481	0.460
γ_7							0.500
$N\mu_N$	0.21	0.42	14.87	0.20	0.29	2.14	2.04
μ_1	5.55×10^{-7}	9.99×10^{-3}	2.38×10^{-4}	3.96×10^{-3}	3.91×10^{-3}	1.52×10^{-3}	1.46×10^{-3}
μ_2		1.47×10^{-6}	3.57×10^{-5}	1.27×10^{-3}	9.37×10^{-4}	5.82×10^{-3}	9.99×10^{-3}
μ_3			1.00×10^{-2}	9.99×10^{-3}	1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}
μ_4				9.98×10^{-3}	6.91×10^{-3}	1.00×10^{-2}	9.99×10^{-3}
μ_5					1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}
μ_6						9.16×10^{-3}	9.99×10^{-3}
μ_7							1.00×10^{-2}

To measure our fitting results, we do the Mann-Whitney test for all fittings of the models with two to eight driver mutations, which shows that all p-values are greater than 0.1. This implies that the models with two to eight driver mutations may be used to analyze the incidence rate of colorectal cancer from SEER and LSS data. To further verify the rationality of fitting results, we discuss the optimal estimated values of parameters for the models with two to eight driver mutations. Tables 3–5 display the simulated optimal values of model parameters. We analyze the number of premalignant cells with different driver mutations at different times by using parameter values in Tables 3–5. The number calculation of premalignant cells can be obtained from Eq (2.4) in the mathematical model. The results are displayed in Figure 4, which shows that the number of premalignant cells in the model with more than five mutations will exceed 10^9 for the SEER data. For LSS data with a low dose level, the premalignant cells of the models with more than four driver mutations will exceed 10^9, and those of the models with more than three driver mutations will exceed 10^9 for the patients with a high dose. However, the environmental maximum capacity of the number of cells in the tissue is usually assumed approximately 10^9 [29]. In addition, it is unlikely for the number of malignant cells to outnumber 10^9 before they spread to other tissues [30,31]. Therefore, the number of premalignant cells should be set to less than 10^9, then, two to five driver mutations are more reasonable to explain the primary colorectal cancer in the SEER data without radiation effect. For LSS data, two to four mutations and two to three mutations are more suitable to analyze the primary colorectal cancer patients with a low dose radiation level and high dose radiation level, respectively. That is, it involves fewer mutations to become malignant cells for the colorectal cancer patients with the dose radiation effect. These imply that radiation exposure can significantly affect the development of colorectal cancer and induce mutation in some key genes that leads to the progress of tumors in colorectal tissue.

Table 5. The net growth rates and mutation rates in the models with two to eight mutations for the LSS data with a high-radiation exposure level ($\gamma_i = a_i - b_i$).

	2	3	4	5	6	7	8
γ_1	0.130	0.040	0.028	0.118	0.119	0.048	0.025
γ_2		0.169	0.190	0.118	0.482	0.118	0.218
γ_3			0.457	0.493	0.495	0.300	0.255
γ_4				0.493	0.496	0.300	0.255
γ_5					0.500	0.335	0.387
γ_6						0.481	0.460
γ_7							0.500
$N\mu_N$	0.27	0.57	13.66	0.23	0.29	2.22	2.14
μ_1	9.68×10^{-7}	1.00×10^{-2}	2.63×10^{-4}	3.89×10^{-3}	4.11×10^{-3}	1.54×10^{-3}	1.48×10^{-3}
μ_2		1.60×10^{-6}	3.77×10^{-5}	1.21×10^{-3}	9.37×10^{-4}	5.82×10^{-3}	9.99×10^{-3}
μ_3			1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}
μ_4				9.99×10^{-3}	6.91×10^{-3}	1.00×10^{-2}	9.99×10^{-3}
μ_5					1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}
μ_6						9.16×10^{-3}	9.99×10^{-3}
μ_7							1.00×10^{-2}

From Tables 3–5, the first mutation rate is extremally small (belongs to the range of $10^{-6} \sim 10^{-8}$) if the number of normal cells, N, equals 10^7 [34,35]. That is, the first event is most likely point mutation [20,21]. The mutation rate of cells with more than three mutations is very high, which may be caused by genetic instability [36–38]. By above analyses, the models with two and three driver mutations are suitable for both the SEER data and the LSS data with different radiation dose levels. Taking the models with two and three mutations as an example, we then analyze the influence of disturbing parameter on the risk of colorectal cancer, which is shown in Figures 5 and 6. The result shows that the parameter γ_2 has a more significant impact in the three-mutation model for SEER data, whereas γ_1 is more sensitive in the two-mutation model for SEER data. In addition, the change of the first proliferation rate, γ_1, has the greatest influence on the risk of colorectal cancer for LSS data with low and high dose levels. By comparing the net growth of mutated cells in Tables 3–5, the net growth rates of cells in patients from LSS data with the radiation exposure effect is higher than those in patents from the SEER data for the models with two to five mutations, which may be caused by radiation dose [39–41]. However, all net growth rates and mutation rates of cells do not always increase with the dose level. In addition, our analysis indicates that the net growth rate of premalignant cells is relatively moderate for the model explaining the colorectal cancer with biological credibility, which is in line with the viewpoint of reference [17].

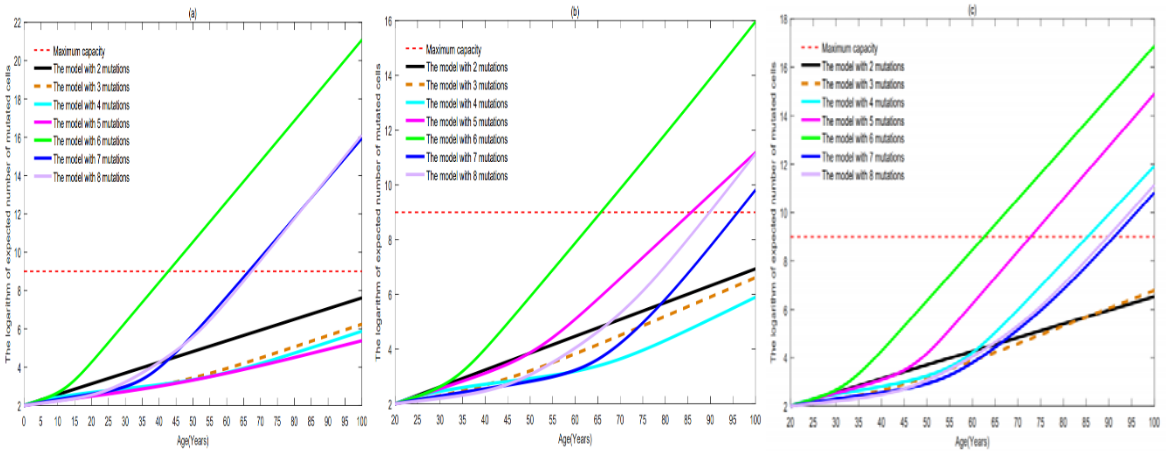

Figure 4. The logarithm of all expected numbers of premalignant cells at different times in the model with different mutation numbers. (a) SEER data. (b) LSS data with low-radiation exposure. (c) LSS data with high-radiation exposure.

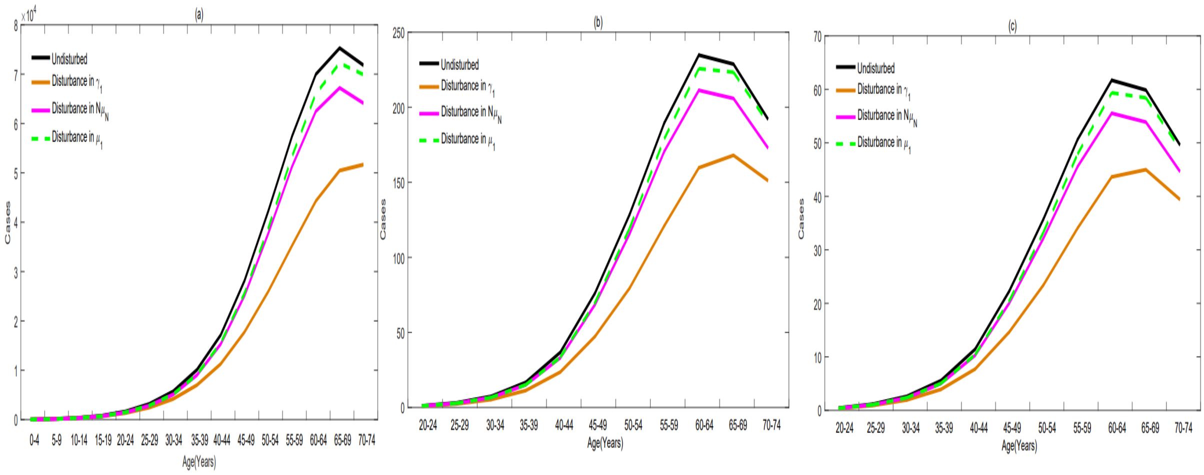

Figure 5. The changes of risk in obtaining colorectal cancer by disturbing parameters of the two-mutation model, and the corresponding parameter value is reduced by a factor of 0.1. (a) SEER data, (b) LSS data with low-radiation exposure, (c) LSS data with high-radiation exposure.

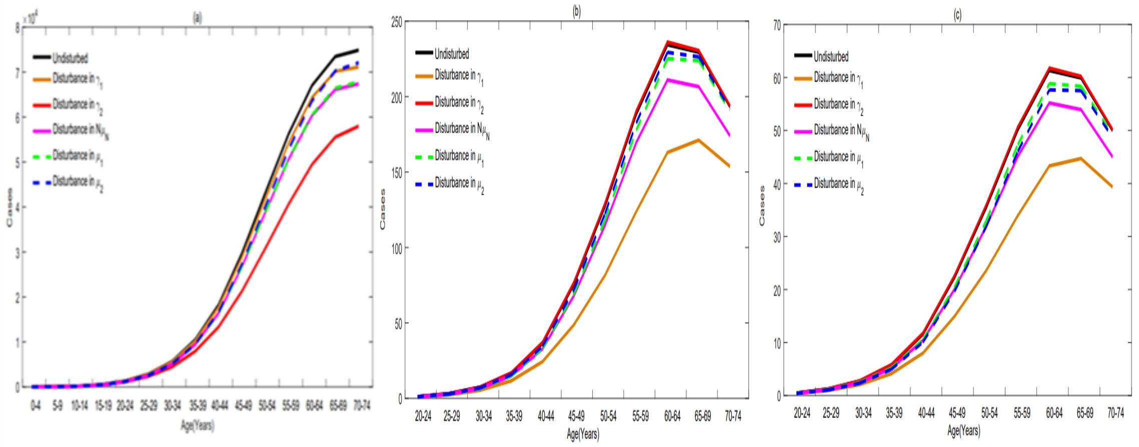

Figure 6. The changes of risk in obtaining colorectal cancer by disturbing parameters of the three-mutation model, and the corresponding parameter value is reduced by a factor of 0.1. (a) SEER data, (b) LSS data with low-radiation exposure, (c) LSS data with high-radiation exposure.

4. Discussion

Stochastic multistage mathematical model has been widely used to study the mechanisms of cancer development. However, most of the existing studies consider the model with a fixed number of gene mutations. It is unreasonable to use one single mathematical model to explain all cancer patients because of the heterogeneity of tumors. Therefore, in our work, the model with any number of gene mutations was allowed to analyze the risk of colorectal cancer. We fit the age-specific incidence rate of colorectal cancer from SEER data and from LSS data with low and high doses levels by using the model with any number of gene mutations. In addition, the fitted optimal parameters analyzed to explain the risk of tumors and the rationality of the fitting results. It is suggested that it requires two to five driver mutations for the colorectal cancer patients without the effect of radiation. However, patients who are exposed to low dose levels of exposure require two to four driver mutations to develop the colorectal cancer, and two to three driver mutations are needed to development a malignant tumor for patients who are exposed to high dose levels of exposure. Furthermore, the proliferation in the last type premalignant cells is more sensitive than that in other premalignant cells for the colorectal cancer patients without radiation effect, and the proliferation rate in mutated cells with one mutation is the most sensitive parameter in the model for the colorectal cancer patients with the radiation effect. The data from LSS involves significant ionizing radiation information. Thus, our result can reflect the effect of radiation exposure on the risk of tumors.

A better understanding of the tumorigenic process will help us to improve our ability to prevent and treat cancer. Our study gave all possible numbers of gene mutations in the development of colorectal cancer for SEER and LSS data involving radiation doses level. This can used to explain the heterogeneity of tumors. It is well known that the activation of oncogene needs one hit and the function loss of the tumor suppressor gene requires two hits [9]. The patients who are not exposed to radiation may undergo 2 to 5 driver mutations. It involves the alternations of two oncogenes or one tumor suppressor gene for two driver mutations, three oncogenes or one tumor suppressor gene and one oncogene for three driver mutations, four oncogenes or one tumor suppressor gene and two oncogenes or two tumor suppressor genes for four driver mutations as well as five oncogenes or one tumor suppressor gene and three oncogenes or two tumor suppressor genes and one oncogene for five driver mutations. These inferences can provide some guidance for the usage of drugs to develop tumor therapy schemes.

Our study only displays all possible numbers of driver mutations for colorectal cancer in the environment without radiation exposure and with low or high doses of radiation exposure. However, it does not consider the other mechanisms through which the radiation effect on the risk of colorectal cancer may occur. Furthermore, the specific oncogene or tumor suppressor gene is still unclear for different numbers of driver mutations in the development of colorectal cancer. These issues will be very valuable for the study of cancer, which will be the direction we need to investigate in the future.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by Natural Science Foundation of China Grant No. 12001417, Postdoctoral Research Foundation of China Grant No. 2023M732168 and Shaanxi Science and Technology Association Young Talent Lifting Program Grant No. 20220519.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA-Cancer J. Clin.*, **71** (2021), 209–249. https://doi.org/10.3322/caac.21660
2. K. L. Newcomer, L. D. Porter, A delayed path to diagnosis: Findings from young-onset colorectal cancer patients and survivors, *J. Clin. Oncol.*, **39** (2021), 5. https://doi.org/10.1200/JCO.2021.39.3_ suppl.5
3. H. J. Li, D. Boakye, X. C. Chen, M. Hoffmeister, H. Brenner, Association of body mass index with risk of early-onset colorectal cancer: Systematic review and meta-analysis, *Am. J. Gastroenterol.*, **116** (2021), 2173–2183. https://doi.org/10.14309/ajg.0000000000001393
4. W. Liu, Y. Deng, Z. Li, Y. Chen, X. Zhu, X. Tan, et al., Cancer evo-dev: a theory of inflammation-induced oncogenesis, *Front. Immunol.*, **12** (2021), 768098. https://doi.org/10.3389/fimmu.2021.768098
5. R. R. Huxley, A. Ansary-Moghaddam, P. Clifton, S. Czernichow, C. L. Parr, M. Woodward, The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence, *Int. J. Cancer*, **125** (2009), 171–180. https://doi.org/10.1002/ijc.24343
6. J. P. Thakkar, B. J. McCarthy, J. L. Villano, Age-specific cancer incidence rates increase through the oldest age groups, *Am. J. Med. Sci.*, **348** (2014), 65–70. https://doi.org/10.1097/maj.0000000000000281

7. S. Jones, W. D. Chen, G. Parmigiani, F. Diehl, N. Beerenwinkel, T. Antal, et al., Comparative lesion sequencing provides insights into tumor evolution, *PNAS*, **105** (2008), 4283–4288. https://doi.org/10.1073/pnas.0712345105
8. L. A. Loeb, Mutator phenotype may be required for multistage carcinogenesis, *Cancer Res.*, **51** (1991), 3075–3079.
9. B. Vogelstein, K. W. Kinzler, Cancer genes and the pathways they control, *Nat. Med.*, **10** (2004), 789–799. https://doi.org/10.1038/nm1087
10. S. Guo, Y. Ye, X. Liu, Y. Gong, M. Xu, L. Song, et al., Intra-tumor heterogeneity of colorectal cancer necessitates the multi-regional sequencing for comprehensive mutational profiling, *Cancer Manag. Res.*, **13** (2021), 9209–9223. https://doi.org/10.2147/cmar.s327596
11. Y. Kamal, G. Idos, Incidental young-onset adenomas: sporadic findings or harbingers of increased colon cancer risk, *Curr. Treat. Options Gastroenterol.*, **20** (2022), 122–132. https://doi.org/10.1007/S11938-022-00375-0
12. V. Wunderlich, Early references to the mutational origin of cancer, *Int. J. Epidemiol.*, **36** (2007), 246–247. https://doi.org/10.1093/ije/dyl272
13. P. Armitage, R. Doll, The age distribution of cancer and a multi-stage theory of carcinogenesis, *Br. J. Cancer*, **8** (1954), 1–12. https://doi.org/10.1038/bjc.1954.1
14. A. G. Knudson, Mutation and cancer: statistical study of retinoblastoma, *PNAS*, **68** (1971), 820–823. https://doi.org/10.1073/pnas.68.4.820
15. E. G. Luebeck, S. H. Moolgavkar, Multistage carcinogenesis and the incidence of colorectal cancer, *PNAS*, **99** (2002), 15095–19100. https://doi.org/10.1073/pnas.222118199
16. R. Meza, J. Jeon, S. H. Moolgavkar, E. G. Luebeck, Age-specific incidence of cancer: phases, transitions, and biological implications, *PNAS*, **105** (2008), 16284–16289. https://doi.org/10.1073/pnas.0801151105
17. E. G. Luebeck, K. Curtius, J. Jeon, W. D. Hazelton, Impact of tumor progression on cancer incidence curves, *Cancer Res.*, **73** (2013), 1086–1096. https://doi.org/10.1158/0008-5472.can-12-2198
18. B. M. Lang, J. Kuipers, B. Misselwitz, N. Beerenwinkel, Predicting colorectal cancer risk from adenoma detection via a two-type branching process model, *PLoS Comput. Biol.*, **16** (2020), e1007552. https://doi.org/10.1371/journal.pcbi.1007552
19. C. Paterson, H. Clevers, I. Bozic, Mathematical model of colorectal cancer initiation, *PNAS*, **117** (2020), 20681–20688. https://doi.org/10.1073/pnas.2003771117
20. A. Niida, K. Mimori, T. Shibata, S. Miyano, Modeling colorectal cancer evolution, *J Hum Genet*, **66** (2021), 869–878. https://doi.org/10.1038/s10038-021-00930-0
21. M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, et al., Discovery and saturation analysis of cancer genes across 21 tumour types, *Nature*, **505** (2014), 495–501. https://doi.org/10.1038/nature12912
22. J. L. Bos, E. R. Fearon, S. R. Hamilton, V. M. Verlaande, J. H. Van-Boom, A. J. Van-der, et al., Prevalence of ras gene-mutations in human colorectal cancers, *Nature*, **327** (1987), 293–297. https://doi.org/10.1038/327293a0
23. E. J. Grant, A. Brenner, H. Sugiyama, R. Sakata, A. Sadakane, M. Utada, et al., Solid cancer incidence among the life span study of atomic bomb survivors: 1958–2009, *Radiat. Res.*, **187** (2017), 513–537. https://doi.org/10.1667/RR14492.1

24. 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** (1988), 383–392. https://doi.org/10.1111/j.1539-6924.1988.tb00502.x
25. C. J. Portier, A. Kopp-Schneider, C. D. Sherman, Calculating tumor incidence rates in stochastic models of carcinogenesis, *Math. Biosci.*, **135** (1996), 129–146. https://doi.org/10.1016/0025-5564(96)00011-9
26. L. Li, T. Tian, X. Zhang, Mutation mechanisms of human breast cancer, *J. Comput. Biol.*, **25** (2018), 396–404. https://doi.org/10.1089/cmb.2017.0111
27. K. S. Crump, R. P. Subramaniam, C. B. Van-Landingham, A numerical solution to the nonhomogeneous two-stage MVK model of cancer, *Risk Anal.*, **25** (2005), 921–926. https://doi.org/10.1111/j.1539-6924.2005.00651.x
28. H. Fakir, W. Y. Tan, L. Hlatky, P. Hahnfeldt, R. K. Sachs, Stochastic population dynamic effects for lung cancer progression, *Radiat. Res.*, **172** (2009), 383–393. https://doi.org/10.1667/rr1621.1
29. R. R. Mercer, M. L. Russell, V. L. Roggli, J. D. Crapo, Cell number and distribution in human and rat airways, *Am. J. Respir. Cell Mol. Biol.*, **10** (1995), 613–624. https://doi.org/10.1165/ajrcmb.10.6.8003339
30. C. Tomasetti, J. Poling, N. J. Roberts, N. R. London, M. E. Pittman, M. C. Haffner, et al., Cell division rates decrease with age, providing a potential explanation for the age-dependent deceleration in cancer incidence, *PNAS*, **116** (2019), 20482–20488. https://doi.org/10.1073/pnas.1905722116
31. C. Simonetto, U. Mansmann, J. C. Kaiser, Shape-specific characterization of colorectal adenoma growth and transition to cancer with stochastic cell-based models, *PLoS Comput. Biol.*, **19** (2023), e1010831. https://doi.org/10.1371/journal.pcbi.1010831
32. D. Peér, S. Ogawa, O. Elhanani, Tumor heterogeneity, *Cancer Cell*, **39** (2021), 1015–1017. https://doi.org/10.1016/j.ccell.2021.07.009
33. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, Cancer genome landscapes, *Science*, **339** (2013), 1546–1558. https://doi.org/10.1126/science.1235122
34. C. Tomasetti, B. Vogelstein, Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions, *Science*, **347** (2015), 78–81. https://doi.org/10.1126/science.1260825
35. C. S. Potten, C. Booth, D. Hargreaves, The small intestine as a model for evaluating adult tissue stem cell drug targets, *Cell Prolif.*, **36** (2003), 115–129. https://doi.org/10.1046/j.1365-2184.2003.00264.x
36. F. Michor, Y. Iwasa, M. A. Nowak, Dynamics of cancer progression, *Nat. Rev. Cancer*, **4** (2004), 197–205.
37. C. J. Kaiser, R. Meckbach, P. Jacob, Genomic instability and radiation risk in molecular pathways to colon cancer, *PLoS One*, **9** (2014), e111024. https://doi.org/10.1371/journal.pone.0111024
38. L. Li, X. Zhang, T. Tian, Mathematical modelling the pathway of genomic instability in lung cancer, *Sci. Rep.*, **9** (2019), 14136. https://doi.org/10.1038/s41598-019-50500-w
39. D. Fernandez-Antoran, G. Piedrafita, K. Murai, S. H. Ong, A. Herms, C. Frezza, et al., Outcompeting p53-mutant cells in the normal esophagus by redox manipulation, *Cell Stem Cell*, **25** (2019), 329–341. https://doi.org/10.1016/j.stem.2019.06.011

40. N. Nori, A hypothesis: radiation carcinogenesis may result from tissue injuries and subsequent recovery processes which can act as tumor promoters and lead to an earlier onset of cancer, *Br. J. Radiol.*, **93** (2020), 20190843. https://doi.org/10.1259/bjr.20190843

41. M. Eidemüller, J. Becker, J. C. Kaiser, A. Ulanowski, A. I. Apostoaei, F. O. Hoffman, Concepts of association between cancer and ionizing radiation: accounting for specific biological mechanisms, *Radiat. Environ. Biophys.*, **62** (2023), 1–15. https://doi.org/10.1007/s00411-022-01012-1