Cancer models and statistical analysis of age distribution of cancers

Alexandr N. Tetearing

March 15, 2021

Abstract

In this paper, mathematical mutational models of the age distribution of cancers are obtained. These are two models – a simple model and a complex model, which takes into account the growth of the cell population and the transmission of mutations to daughter cells. Using the resulting formulas, we approximated real age-specific cancer incidence datasets in women (colon, lung, mammary, stomach) and men (colon, lung, prostate, stomach). We estimated parameters such as the average number of mutations (per cell per unit of time) and number of mutations required for cancer to occur.

The number of mutations averaged (over four types of cancer) required for cancer to occur is 72 (mutations per cell for women) and 221 (mutations per cell for men) for the complex mutational model. The average (for four types of cancer) mutation rate is estimated at 0.05 mutations per year (per cell for women) and at 0.07 mutations per year (per cell for men).

The mutational models do not provide a quantitative explanation for the occurrence of cancer due to cellular mutations.

As an alternative to mutational models, a model of a single carcinogenic event with an age shift is considered. The model approximates real data better and explains quantitatively the real age distributions of cancers.

Contents

1. Introduction **2**

I Cancer models **3**

2. Simple mutational model **3**
2.1. Mathematical model 3
2.2. Age distribution of cancer incidence 11

3. Complex mutational model **13**
3.1. Mathematical model 13
3.2. Age distribution of cancer incidence 19

4. Model of Delayed Carcinogenic Action **21**

II Statistical analysis of age distributions of cancers **24**

5. Practical datasets **24**

6. Statistical analysis based on a simple mutation model **24**

7. Statistical analysis based on a complex mutation model **27**

8. Statistical analysis based on model of delayed carcinogenic action **31**

III Discussion **34**

9. Mutational models **34**

10. Model of Delayed Carcinogenic Action **36**

1. Introduction

One of the hallmarks of a cancer cell is a large number of hereditary mutations. In the work of Lawrence Loeb [1], the number of mutations is estimated at a huge number of 10,000 - 100,000 per cell. But this number only applies to long-lived progressive tumors, in which most of the mutations could arise after the degeneration of a normal cell into a cancerous one.

But how many mutations are the minimum required for the formation of a cancer cell?

Alfred Knudson was the first to try to answer this question in 1971. Knudson investigated retinoblastoma, an ocular tumor that occurs in childhood [2]. Based on the available data on 48 cases of retinoblastoma, Using the Poisson distribution (mathematical probability distribution of a random event), Knudson concluded that the average number of mutations leading to cancer is three mutations per cell. And the first mutation is a hereditary mutation passed on to a child from a parent.

For adult forms of cancer, the question of the minimum required number of cell mutations leading to cancerous degeneration remains open.

In this work, mathematical formulas are obtained that describe the age distribution of cancer in a group of patients or in a cell population. When changing the parameter z in the formula (this is the number of mutations required for a cancerous transformation of a cell), the curve describing the age distribution of cancer in the population, changes its shape. Comparing theoretical curves with real (obtained in practice) age distributions, we can choose the value of z parameter so that the theoretical distribution most closely matches the practical age distribution of cancer.

Then we will know approximately the average minimum number of mutations required for a cancerous cell transformation (if the used mathematical model accurately describes the process of mutation formation).

Part I
Cancer models

2. Simple mutational model

2.1. Mathematical model

In the simplest model, we consider a constant cell set (population of cells), consisting of N_0 individual cells. Let N_0 be 1000, and the number of cells does not change over time. All cells are constantly exposed to destructive external influence. As a result of external influence, cells receive mutations. The average number of mutations in a cell per unit time is constant and equal to a. The average number of mutations in a cell in a short time dt is the product of a and dt. The average number of mutations in a cell population of N_0 cells in a short time dt is $N_0\,a\,dt$.

We will assume that, with a certain number of mutations accumulated by the cell (we will denote this number as z), the cell (in any case) is transformed

into a cancer cell (perhaps this does not happen immediately, but only after a long time).

Our primary task is to calculate the number of cells in a population with different numbers of mutations.

The damaging effect on the cell population begins at the time $t = 0$. At the time $t = 0$ all cells do not have mutations, and the number of cells with mutations is zero. At the time $t = dt$ the number of mutated cells $N1$ is:

$$N1(dt) = N_0\, a\, dt \tag{2.1}$$

The number of cells without mutations $N(t)$ is:

$$N(dt) = N_0\,(1 - a\, dt) \tag{2.2}$$

At the time $t = 2dt$, the number of normal (non-mutated) cells N is:

$$N(2dt) = N_0\,(1 - a\, dt) - N_0\,(1 - a\, dt)\, a\, dt = N_0\,(1 - a\, dt)^2 \tag{2.3}$$

At the time $t = 3dt$, the number of normal (non-mutated) cells N is:

$$N(3dt) = N_0\,(1 - a\, dt)^3, \tag{2.4}$$

etc.

Thus, the number of normal cells at the time $t = T$ is the multiplical of the negative constant a on the interval $[0; T]$:

$$N(T) = N_0 \prod_0^T (-a)\, dt \tag{2.5}$$

The definite multiplical $\prod_A^B$ of the function $f(t)$ is the limit of the infinite product:

$$\prod_A^B f(t)\, dt = \lim_{n\to\infty} \prod_{k=1}^{n} \left[1 + f\left(A + k\, \frac{(B - A)}{n} \right) \frac{B - A}{n} \right] \tag{2.6}$$

The basic formula of the multiplical calculus expresses the definite multiplical of the function $f(t)$ as the difference between the values of the antiderivative $F(t)$ at the ends of the interval:

$$\prod_A^B f(t)\, dt = exp\Big(F(B) - F(A) \Big), \tag{2.7}$$

$$F(t) = \int f(t)\, dt$$

The formula (2.7) is given in [3]. In our case, since a is a constant:

$$N(T) = N_0 \int_0^T (-a)\, dt = e^{-aT} \tag{2.8}$$

The number of cells with one mutation (or more) at time t is:

$$N1(t) = N_0 (1 - e^{-at}) \tag{2.9}$$

The graph of the function $N1(t)$ is shown in Fig. 2.1.

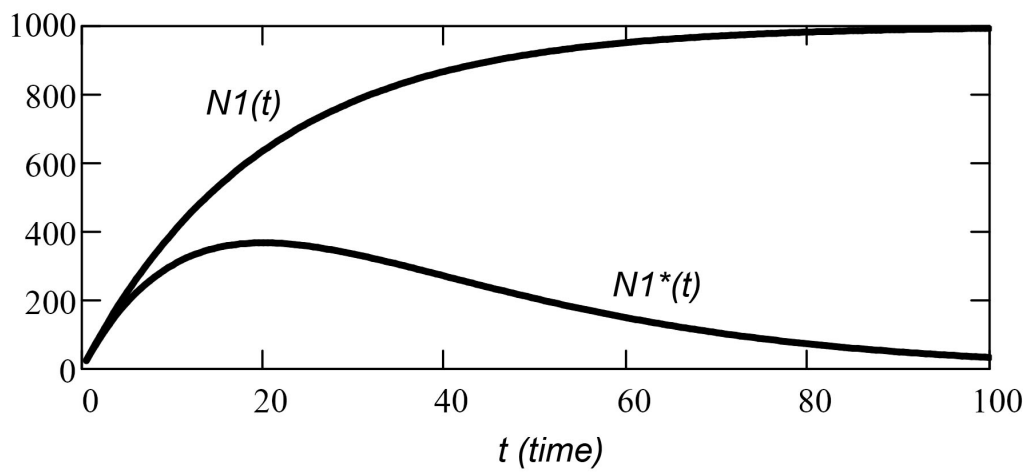

Рис. 2.1. The number of cells with one (or more) mutations is the function $N1(t)$ (upper curve in the figure). The number of cells with one mutation (at most one) is the function $N1^*(t)$ (bottom curve).

The following parameters were used to plot the graph: $N_0 = 1000$; $a = 0.05$.

Now let us find the number $N1^*(t)$ of cells in the population that at the time $t = T$ have just one mutation (no more, no less).

Cells with one mutation are formed from cells without mutations, the number of which is described by the function N (t). In each time interval dt (starting from the time $t = dt$) there are two processes that change the number of cells with one mutation. On the one hand, the available number of cells with one mutation decreases due to the fact that some of the cells receive the second mutation. On the other hand, the number of cells with one mutation increases due to the fact that some cells without mutations receive one mutation (on each time interval dt this is the increment of the function $N1(t)$ in time dt).

Since the function $N1(t)$ is a function that changes over time, we cannot use the formula in the form (2.5) to find $N1^*(t)$. We have to calculate the

multiplical of increment of the function $N1(t)$ (the multiplical of $dN1(t)$) for each interval dt instead.

Let's look at Fig. 2.2.

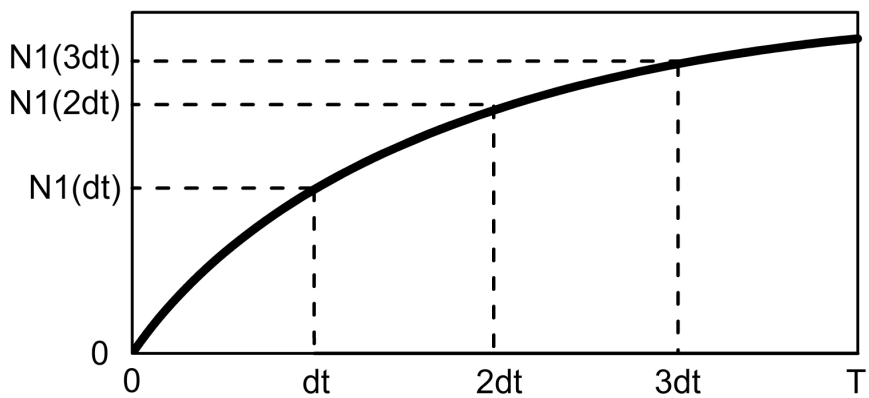

Рис. 2.2. Calculation of the function $N2(t)$ (the number of cells with two mutations).

The function $N1(t)$ is bounded, positive, and monotonically increasing on the interval $[0;T]$.

We divide the interval $[0;T]$ into small segments dt. The number of these segments is $n = T/dt$. In this case, $n = 4$. Then we can find the function $N1^*(t)$ as the sum:

$$N1^*(T) = N1(dt)\int_{dt}^{T}(-a)\,dt + \Big(N1(2dt) - N1(dt)\Big)\int_{2dt}^{T}(-a)\,dt + \\ + \Big(N1(3dt) - N1(2dt)\Big)\int_{3dt}^{T}(-a)\,dt + \Big(N1(4dt) - N1(3dt)\Big) \qquad (2.10)$$

or:

$$N1^*(T) = f(0)\,dt\,e^{-a\,(T-dt)} + f(dt)\,dt\,e^{-a\,(T-2dt)} + \\ + f(2dt)\,dt\,e^{-a\,(T-3dt)} + f(3dt)\,dt, \qquad (2.11)$$

where $f(t)$ is the time derivative of the function $N1(t)$:

$$f(t) = N1`(t) = \frac{d\Big(N1(t)\Big)}{dt} \qquad (2.12)$$

If dt is infinitely small (with $n \longrightarrow \infty$), then it is true:

$$f(t) \overset{dt\to 0}{\approx} f(t+dt) \qquad (2.13)$$

And $f(0) \overset{dt\to 0}{\longrightarrow} f(dt)$, $f(dt) \overset{dt\to 0}{\longrightarrow} f(2dt)$, etc. Moreover, $f((n-1)dt)\,dt \overset{dt\to 0}{\longrightarrow} 0$. Then:

$$N1^*(T) = e^{-aT} \int_0^T N1'(t)\, e^{at} dt \tag{2.14}$$

We will use this formula in the future, but, in this case:

$$N1^*(T) = e^{-aT} \int_0^T N_0\, a\, e^{-a\,t}\, e^{at} dt = e^{-aT}\, N_0\, a \int_0^T dt \tag{2.15}$$

Thus, the number of cells with one mutation changes over time as a function of $N1^*(t)$:

$$N1^*(t) = N_0\, a\, t\, e^{-a\,t} \tag{2.16}$$

There is another way to calculate the function $N1^*(t)$ (the number of cells with one mutation). To do this, we need to solve the differential equation:

$$dN1^*(t) = dN1(t) - N1^*(t)\, a\, dt \tag{2.17}$$

The left side of the equation is the change in the number of cells with one mutation (on the time interval dt). The first term on the right side is the number of cells that received one mutation in dt time (these cells are formed from cells that have not previously been mutated). The second (negative) term on the right is the number of cells that received the second mutation in the dt interval (these cells are formed from cells that previously had only one mutation).

Dividing the equation by dt we have:

$$\frac{dN1^*(t)}{dt} = \frac{dN1(t)}{dt} - N1^*(t)\, a \tag{2.18}$$

We will use the notation:

$$y = N1^*(t); \quad y' = \frac{dN1^*(t)}{dt}; \quad f' = \frac{dN1(t)}{dt} \tag{2.19}$$

Linear non-homogeneous differential equation:

$$y' + a\, y = f' \tag{2.20}$$

is solved by the substitution $y = uv$.

$$u'v + v'u + a\, uv = f' \Rightarrow u'v + u\,(v' + a\, v) = f' \tag{2.21}$$

If the expression in parentheses is zero: $(v' + a\,v) = 0$, we get the following expression for v:

$$(v' + a\,v) = 0 \;\Rightarrow\; v' = -a\,v \;\Rightarrow\; v = e^{-a\,t} \tag{2.22}$$

The resulting expression is substituted into equation (2.21):

$$u' e^{-a\,t} = f' \Rightarrow\; u' = f' e^{a\,t} \tag{2.23}$$

Integrating both sides of the equation, we get:

$$u = c_1 + \int f' e^{a\,t} dt \tag{2.24}$$

Then we substitute the functions u and v into the expression $y = uv$ and return to the notation (2.19):

$$N1^*(t) = c_1 e^{-a\,t} + e^{-a\,t} \int \frac{dN1(t)}{dt}\, e^{a\,t} dt \tag{2.25}$$

Substitution the function $N1(t)$ into this equation gives (see Eq. 2.9):

$$N1^*(t) = c_1 e^{-a\,t} + e^{-a\,t} \int N_0\, a\, e^{-a\,t}\, e^{a\,t} dt \tag{2.26}$$

$$N1^*(t) = c_1 e^{-a\,t} + N_0\, a\, t\, e^{-a\,t} \tag{2.27}$$

To find the integration constant c_1, we substitute $t = 0$ into this equation:

$$c_1 = N1^*(0) \tag{2.28}$$

In our case, $N1^*(0) = 0$. As a result, we get:

$$N1^*(t) = N_0\, a\, t\, e^{-a\,t} \tag{2.29}$$

Thus, we got the equation (2.16) in a different way.

Number of cells with two (or more) mutations $N2$ is a difference of the functions $N1(t)$ and $N1^*(t)$:

$$N2(t) = N1(t) - N1^*(t) = N_0\,(1 - e^{-a\,t} - a\,t\,e^{-a\,t}) \tag{2.30}$$

The graph of the function $N2(t)$ is shown in Fig. 2.3 (upper curve).

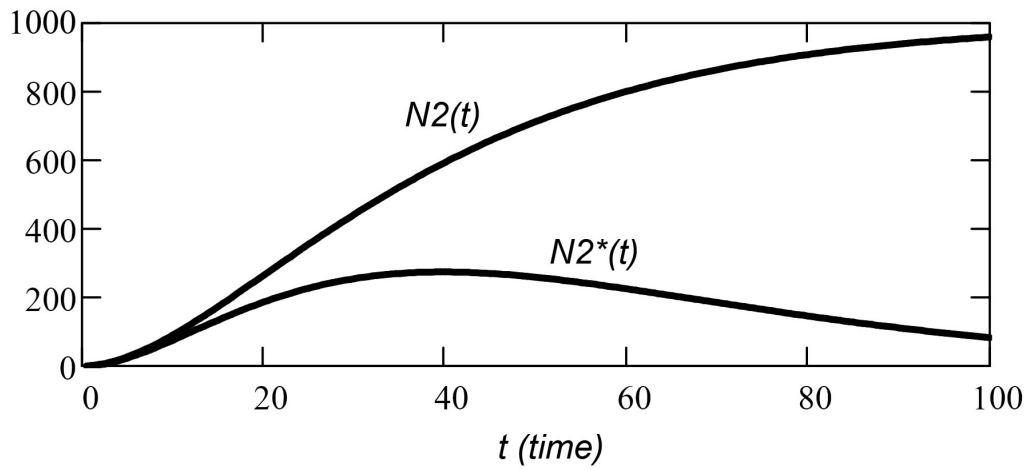

Рис. 2.3. The number of cells with two (or more) mutations (function $N2(t)$ and the upper curve in the figure). The number of cells with exactly two (no more, no less) mutations (function $N2^*(t)$ and the bottom curve).

To find $N2^*$ the number of cells with only two mutations (no more, no less), we will use the formula (2.14) obtained earlier. We need to substitute the function $N2(t)$ into this formula (instead of the function $N1(t)$ as before).

$$N2^*(T) = e^{-aT} \int_0^T N2'(t)\, e^{at} dt = N_0\, a^2\, e^{-aT} \int_0^T t\, dt$$

$$N2^*(t) = N_0\, \frac{(a\,t)^2}{2} e^{-at} \qquad (2.31)$$

The graph of the function $N2^*(t)$ is shown in Fig. 2.3 (bottom curve).

If we subtract the function $N2^*(t)$ (number of cells with exactly two mutations) from the function $N2(t)$ (number of cells with two or more mutations), we get the function $N3(t)$ (the number of cells with three or more mutations).

$$N3(t) = N_0 \left(1 - e^{-at} - a\,t\,e^{-at} - \frac{(a\,t)^2}{2} e^{-at}\right) \qquad (2.32)$$

To find the function $N3(t)$ (number of cells with exactly three mutations), we again use formula (2.14).

In this time we need to substitute the function $N3(t)$ into the formula instead of the function $N1(t)$:

$$N3^*(T) = e^{-aT} \int_0^T N3'(t)\, e^{at} dt = e^{-aT} N_0\, \frac{a^3}{2} \int_0^T t^2 dt$$

$$N3^*(t) = e^{-at} N_0 \frac{(a\,t)^3}{6} \tag{2.33}$$

The graphs of the functions $N3(t)$ and $N3^*(t)$ are shown in Fig. 2.4.

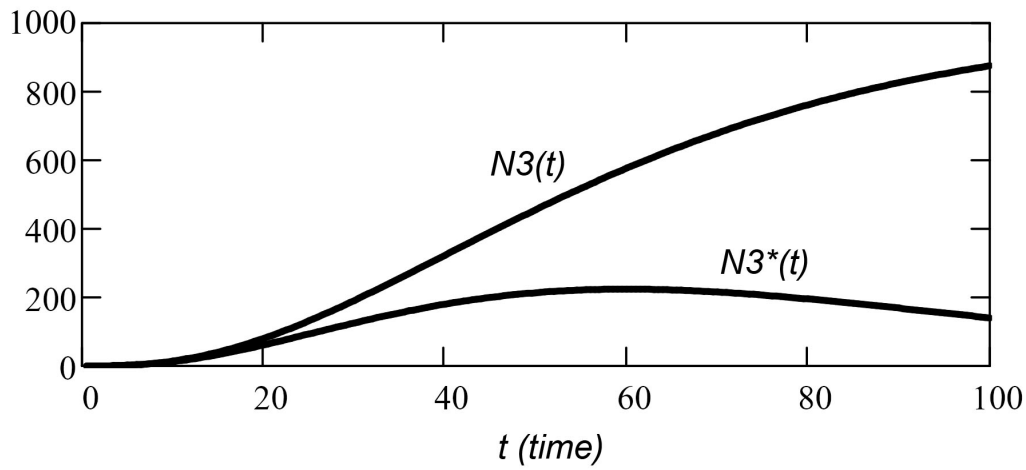

Рис. 2.4. Number of cells with three (or more) mutations (the function $N3(t)$ and the upper curve in the figure). Number of cells with exactly three mutations (the function $N3^*(t)$ and the bottom curve).

Similarly, we find the functions $N4(t) = N3(t) - N3^*(t)$ (number of cells with four or more mutations) and $N4^*(t)$ (number of cells with exactly four mutations).

$$N4(t) = N_0\left(1 - e^{-at}\left(1 + a\,t + \frac{(a\,t)^2}{2} + \frac{(a\,t)^3}{6}\right)\right) \tag{2.34}$$

To find the function $N4^*(t)$ (number of cells with exactly four mutations) we use formula (2.14). We substitute the function $N4(t)$ into the formula instead of the function $N1(t)$:

$$N4^*(T) = e^{-aT} \int_0^T N4'(t)\, e^{at} dt = \tag{2.35}$$

$$N4^*(t) = e^{-at} N_0 \frac{(a\,t)^4}{24} \tag{2.36}$$

The graphs of the functions $N4(t)$ and $N4^*(t)$ are shown in Fig. 2.5.

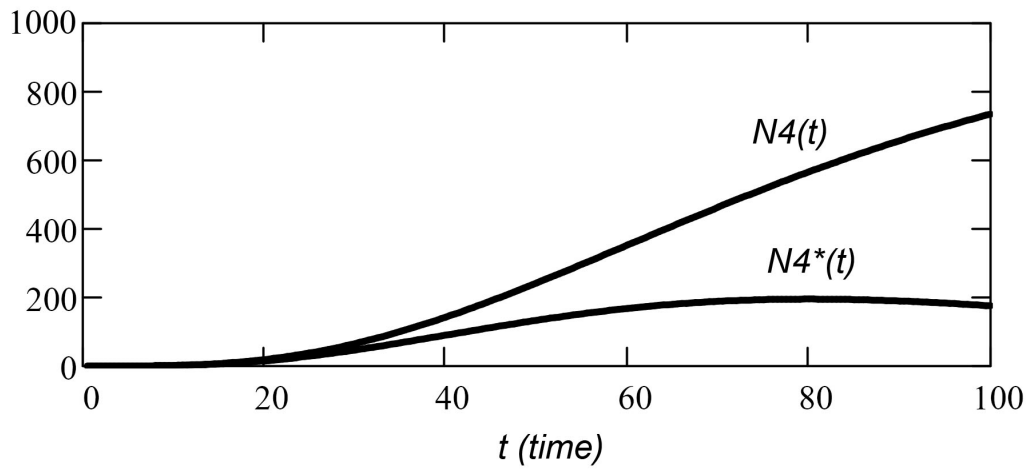

Рис. 2.5. Number of cells with four (or more) mutations (the function $N4(t)$ and the upper curve in the figure). Number of cells with exactly four mutations (the function $N4^*(t)$ and the bottom curve).

Looking at the formulas for the functions $N1(t)$, $N2(t)$, $N3(t)$ and $N4(t)$, we can derive the general formula for the function $Nz(t)$, where z is the minimum number of mutations in the cell. For example, $N8(t)$ is a function that describes the number of cells that have 8 or more mutations.

$$Nz\,(t) = N_0 \left(1 - e^{-at} \sum_{k=0}^{z-1} \frac{(at)^z}{z!}\right)$$

2.2. Age distribution of cancer incidence

We recall that at the beginning of this section we agreed (this is just for example) that if a cell gets four (or more) mutations, the cell becomes cancerous. To find the age distribution of cancers in the cell population, we need to calculate the derivative of the function $Nz(t)$ (in our case, this is the function $N4(t)$). The function increment on an time interval dt is the number of cancers that appeared during this time period.

For example, the age distribution of cancers for the four mutations $D4(t)$ is:

$$D4(t) = \frac{d\big(N4(t)\big)}{dt} = N_0\, e^{-at}\, \frac{a^4\, t^3}{6}$$

The distribution is shown in Fig. 2.6.

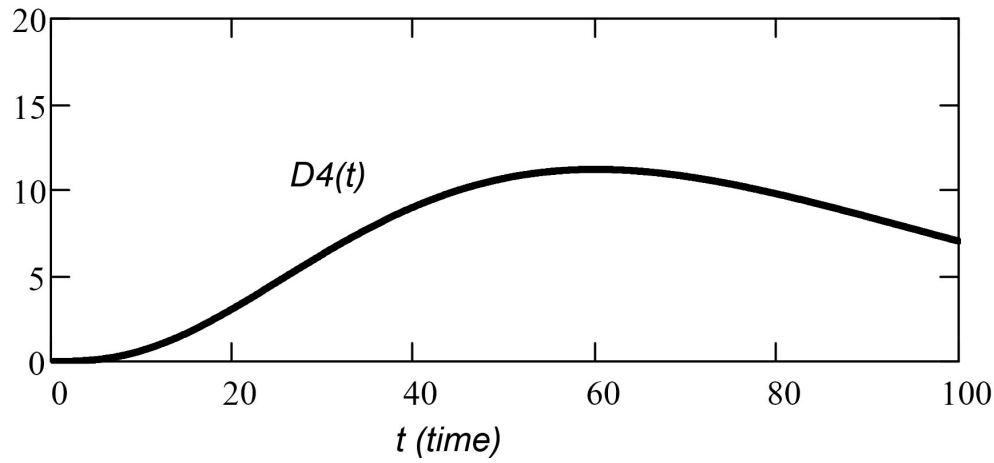

Рис. 2.6. Age distribution of cancers in the cell population. The case when the disease begins with four mutations in the cell.

The general formula for the distribution of cancers that arise as a result of a certain number of mutations in a cell is as follows:

$$Dz(t) = N_0\, e^{-at}\, a\, \frac{(a\,t)^{(z-1)}}{(z-1)!}\,, \qquad (2.37)$$

where z is the number of mutations in the cell, after which cancer necessarily begins.

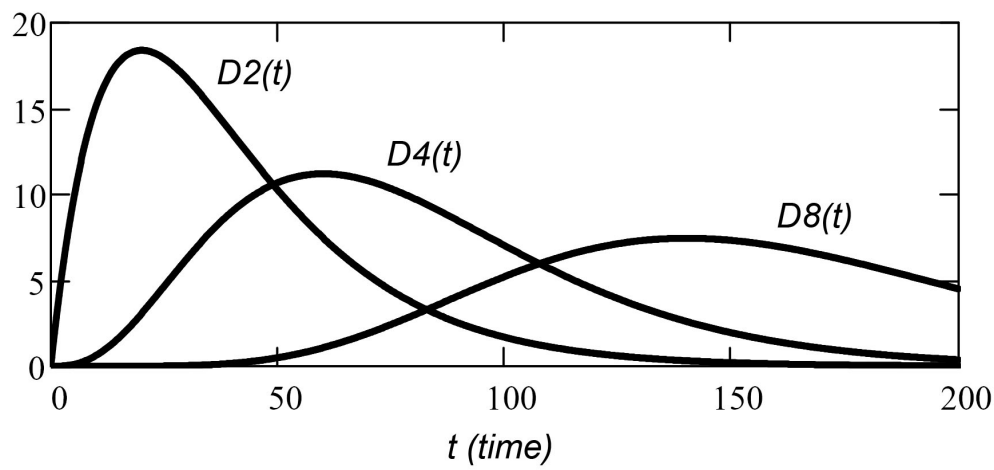

Рис. 2.7. Age distribution of cancers in the cell population. The cases when the disease begins with two mutations in the cell (the function $D2(t)$), with four mutations in the cell (the function $D4(t)$), and with eight mutations in the cell (the function $D8(t)$).

This distribution is known in mathematics as the Erlang distribution [4]. This distribution was obtained by the Danish mathematician Agner Krarup Erlang when he was working on the problem of determining the waiting time for a response from a telephone exchange operator.

The graphs of the age distribution for different numbers of mutations are shown in Fig. 2.7. These are the cases when the disease begins with two mutations in the cell (the function $D2(t)$), with four mutations in the cell (the function $D4(t)$), and with eight mutations in the cell (the function $D8(t)$). For all considered functions, the initial number of cells in the population is $N_0 = 1000$, the average number of mutations per cell per unit of time is $a = 0.05$.

The number of cancers that occur in the population in the interval $[t; t+dt]$, is the product of the value of the age distribution function $Dz(t)$ and the length of the interval dt.

3. Complex mutational model

3.1. Mathematical model

In the previous section of the article, we considered a cell population with a constant number of cells. However, this model does not fit reality well. Real cell tissue grows, increasing the number of cells by several orders of magnitude during the growth process. And cells that have the mutation pass the mutation on to their offspring.

Consider a cell population consisting of one cell $N(0) = 1$ at the initial moment of time $t = 0$. Here $N(0)$ is the number of cells at time $t = 0$.

The growth of the cell population follows the equations:

$$N(t) = \begin{cases} e^{bt} & \text{if } t \le t_1 \\ \dfrac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t} & \text{if } t > t_1 \end{cases} \tag{3.1}$$

Until the time $t = t_1$, the function is a solution to the equation of unlimited growth – the cell population consumes as much food (energy) resource as it can absorb:

$$\frac{dM}{dt} = \frac{k_1 - k_0}{k_3} M(t) \tag{3.2}$$

Here $M(t)$ is the mass of the cell population. To find the size of the population $N(t)$, we need to divide the mass of the population by the average mass of one cell m_0:

$$N(t) = \frac{M(t)}{m_0}$$

In our case, we will assume that the cell mass is equal to one.

After time $t = t_1$, the mass $M(t)$ of the cell population is a solution to the equation of limited growth – the population consumes a limited (constant) amount of food (energy) resource, because the environment cannot provide more resources:

$$\frac{dM}{dt} = \frac{R_0}{k_3} - \frac{k_0}{k_3} M(t) \tag{3.3}$$

The R_0 constant is the amount of food resource that is available to the population (the amount of resource that the environment provides per unit of time).

The constant b from the upper equation of the system (3.1) is determined by the expression:

$$b = \frac{k_1 - k_0}{k_3} \tag{3.4}$$

The time t_1 (the time when the cell population goes from the stage of unlimited growth to the stage with limited resource consumption) is determined by the equality:

$$e^{b\,t_1} = \frac{R_0}{k_1} \tag{3.5}$$

This equality arises from the condition:

$$k_1\, M(t)\, dt = R_0\, dt \tag{3.6}$$

(See. Eq. 3.2 and 3.3).

Under this condition, equation (3.2) turns into equation (3.3).

To find the constant c_3 (this is the constant of integration) we solve the equation:

$$M(t_1) = \frac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3} t_1} \tag{3.7}$$

(see System of Equations 3.1).

All the above equations and coefficients characterizing the cell population are described in detail in [5]. The development graph of the cell population is shown in Fig. 3.1 (the function $N(t)$ is the number of cells in the population).

The $N(t)$ function (this is number of cells in a population) demonstrates exponential growth up to the time $t = t_1$ and limited growth after the time $t = t_1$. The population size at the time $t = t_1$ is $N(t_1) = N_R$. With limited growth, the number of cells gradually reaches an asymptotic plateau $N(\infty) = 1000$ over time.

For plotting graph 3.1, we used the following parameters: $R_0 = 100$; $k_0 = 0.1$; $k_1 = 0.2$; $k_3 = 0.4$.

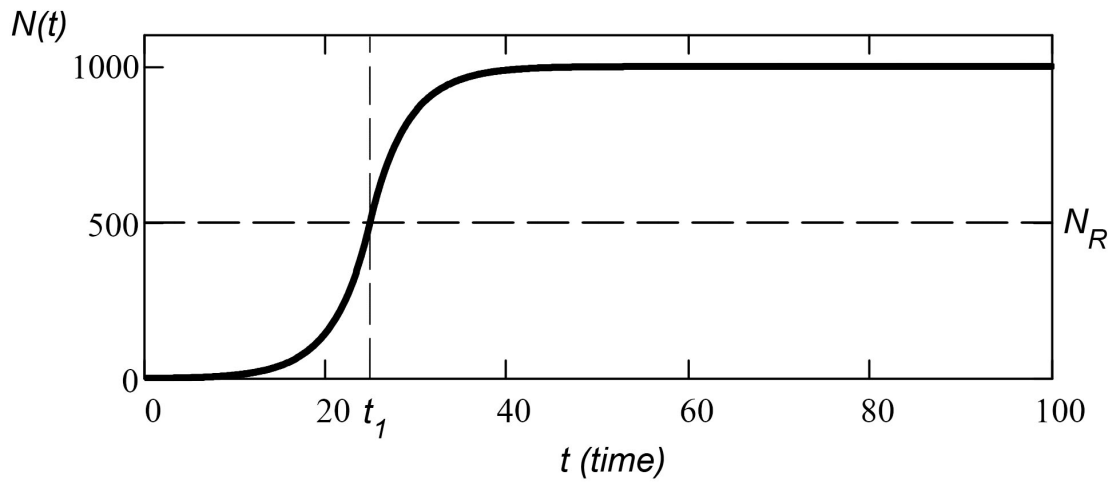

Рис. 3.1. Function $N(t)$ is the number of cells in the population. The cell population has exponential growth until time $t = t_1$ and limited growth since time t_1.

The number of cells that do not have mutations $N0(t)$ is the product of the function $N0(t)$ and the definite multiplical of a constant negative number a on the interval $[0; T]$ (a is the average number of mutations per cell per unit of time):

$$N0(T) = N(T) \int_0^T (-a)\, dt, = N(T)\, e^{-aT} \tag{3.8}$$

In our case:

$$N0(t) = \begin{cases} e^{t(b-a)} & \text{if } t \le t_1 \\ \left(\dfrac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t}\right) e^{-at} & \text{if } t > t_1 \end{cases} \tag{3.9}$$

The number of cells with one mutation (or more) $N1(t)$ is the difference of the function $N(t)$ and $N0(t)$:

$$N1(T) = N(t) - N0(t) \tag{3.10}$$

The number of cells with exactly one mutation $N1^*(t)$ (no less, no more) is the product of the function $N1(t)$ and the definite multiplical of a constant negative number a:

$$N1^*(T) = N1(T) \int_0^T (-a)\, dt \tag{3.11}$$

Here T and t are the same variable. Different notations are used to avoid confusion with the limits of the multiplical.

In our case:

$$N1^*(t) = \begin{cases} e^{t(b-a)}\left(1 - e^{-at}\right) & \text{if } t \le t_1 \\ e^{-at}\left(1 - e^{-at}\right)\left(\dfrac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t}\right) & \text{if } t > t_1 \end{cases} \tag{3.12}$$

The graphs of the functions $N1(t)$ (number of cells with one or more mutations) and $N1^*(t)$ (number of cells with exactly one mutation) are shown in Fig. 3.2.

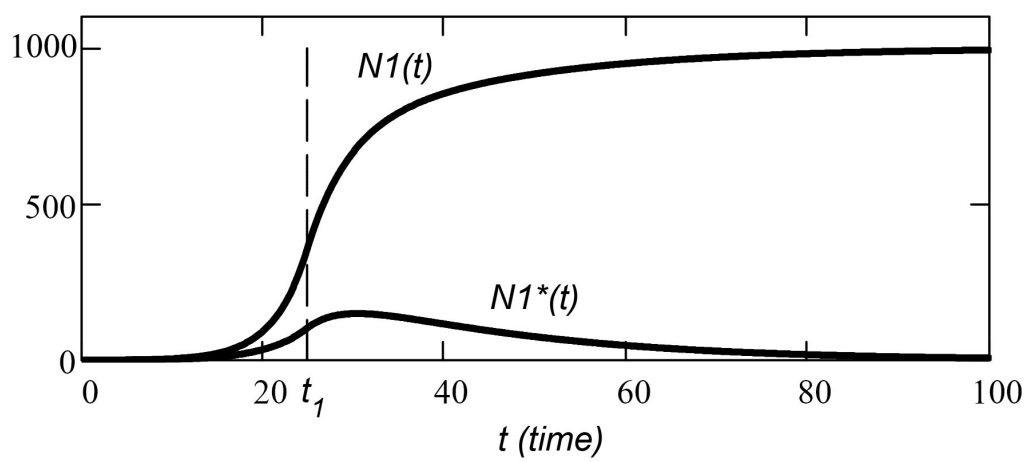

Рис. 3.2. Number of cells with one or more mutations (the function $N1(t)$ and upper curve in the figure) and number of cells with exactly one mutation (the function $N1^*(t)$ and bottom curve).

To find the number of cells with two or more mutations (function $N2(t)$), we need to subtract the function $N1^*(t)$ from the function $N1(t)$:

$$N2(t) = N1(t) - N1^*(t)$$

Or:

$$N2(t) = \begin{cases} e^{bt}\left(1 - e^{-at}\right)^2 & \text{if } t \le t_1 \\ \left(1 - e^{-at}\right)^2\left(\dfrac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t}\right) & \text{if } t > t_1 \end{cases} \tag{3.13}$$

The number of cells with exactly two mutations (the function $N2^*(t)$) is:

$$N2^*(T) = N2(T) \int_0^T (-a)\, dt$$

$$N2^*(t) = \begin{cases} e^{-a\,t}\, e^{b\,t} \left(1 - e^{-a\,t}\right)^2 & \text{if } t \le t_1 \\ e^{-a\,t} \left(1 - e^{-a\,t}\right)^2 \left(\dfrac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3} t}\right) & \text{if } t > t_1 \end{cases} \qquad (3.14)$$

The graphs of the functions $N2(t)$ (the number of cells with two or more mutations) and $N2^*(t)$ (the number of cells with exactly two mutations) are shown in Fig. 3.3.

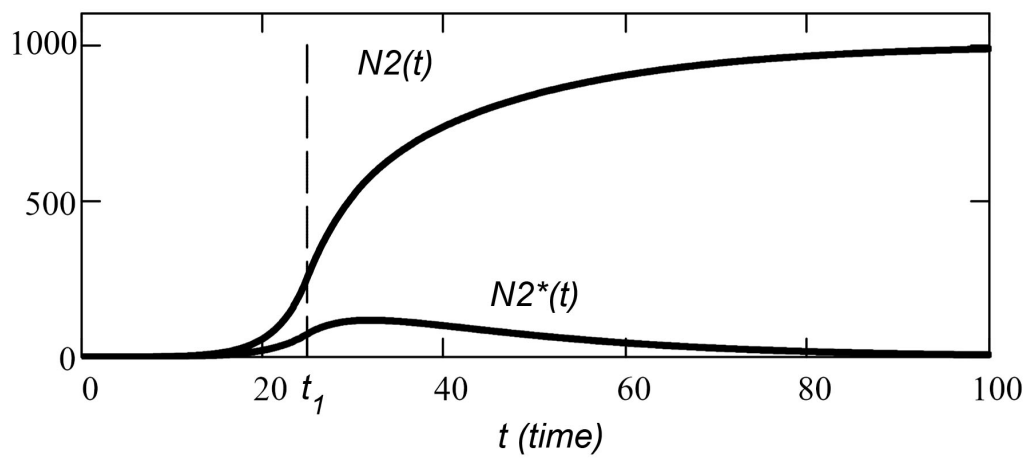

Рис. 3.3. Number of cells with two or more mutations (the function $N2(t)$ and upper curve in the figure) and number of cells with exactly two mutations (the function $N2^*(t)$ and bottom curve).

The formulas for the function $N3(t)$ (the number of cells with three or more mutations) are derived in a similar way.

$$N3\,(t) = N2(t) - N2^*\,(t)$$

Or:

$$N3\,(t) = \begin{cases} e^{b\,t} \left(1 - e^{-a\,t}\right)^3 & \text{if } t \le t_1 \\ \left(1 - e^{-a\,t}\right)^3 \left(\dfrac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3} t}\right) & \text{if } t > t_1 \end{cases} \qquad (3.15)$$

The function $N3^*(t)$ (the number of cells with exactly three mutations) is:

$$N3^*(T) = N3(T) \int_0^T (-a)\, dt)$$

$$N3^*(t) = \begin{cases} e^{-a\,t}\,e^{b\,t}\left(1-e^{-a\,t}\right)^3 & \text{if } t \le t_1 \\ e^{-a\,t}\left(1-e^{-a\,t}\right)^3\left(\dfrac{R_0}{k_0} - c_3\,e^{-\frac{k_0}{k_3}t}\right) & \text{if } t > t_1 \end{cases} \tag{3.16}$$

The graphs of the functions $N3(t)$ (the number of cells with three or more mutations) and $N3^*(t)$ (the number of cells with exactly three mutations) are shown in Fig. 3.4.

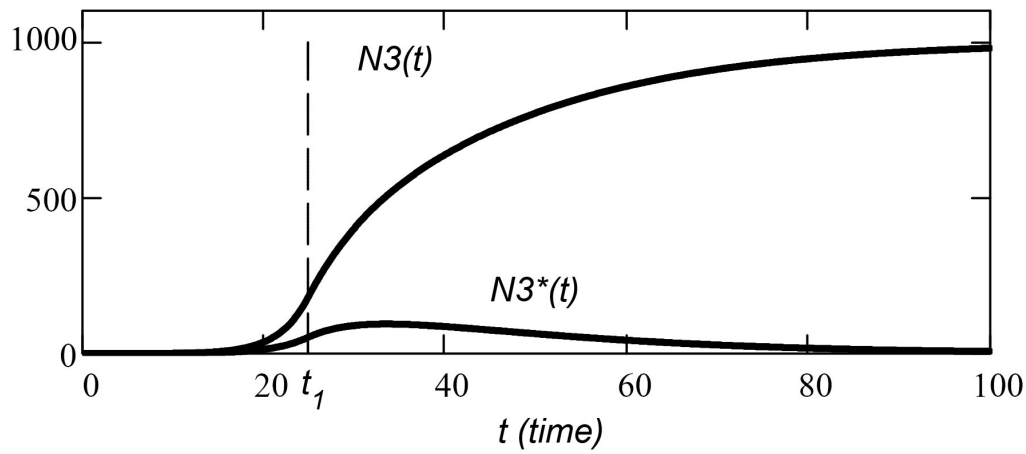

Рис. 3.4. Number of cells with three or more mutations (the function $N3(t)$ and upper curve in the figure) and number of cells with exactly three mutations (the function $N3^*(t)$ and bottom curve).

Looking at these functions, we can derive a general formula for the function $Nz(t)$ with a certain (minimum specified) number of mutations (where z is the number of mutations in the cell):

$$Nz(t) = \begin{cases} e^{b\,t}\left(1-e^{-a\,t}\right)^z & \text{if } t \le t_1 \\ \left(1-e^{-a\,t}\right)^z\left(\dfrac{R_0}{k_0} - c_3\,e^{-\frac{k_0}{k_3}t}\right) & \text{if } t > t_1 \end{cases} \tag{3.17}$$

The function $Nz(t)$ describes the number of cells in the population in which the number of mutations is equal to or exceeds z. We recall that the time $t = t_1$ is the initial parameter of the cell population, which is determined by formula (3.5).

3.2. Age distribution of cancer incidence

In order to obtain the age distribution $Dz(t)$ of cancer incidence (in the cell population), we need to find the time derivative of the function $Nz(t)$:

$$Dz(t) = \frac{d\left(Nz(t)\right)}{dt} \tag{3.18}$$

For example, the age distribution $D3(t)$ for the function $N3(t)$ (the disease begins with three or more mutations in the cell) is found by the formula:

$$D3(t) = \frac{d\left(N3(t)\right)}{dt}$$

$$D3(t) = \begin{cases} e^{b\,t}\left(1 - e^{-a\,t}\right)^{z} & \text{if } t \le t_1 \\ \left(1 - e^{-a\,t}\right)^{z}\left(\frac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t}\right) & \text{if } t > t_1 \end{cases} \tag{3.19}$$

To find the age distribution for a group of N_P people, we need to multiply the resulting formula by N_P. Thus, the general formula for the age distribution of cancer incidence (the function $Dz(t)$) in a group of N_P people is:

$$Dz(t) = \begin{cases} N_P\, e^{b\,t}\left[b\left(1 - e^{-a\,t}\right)^{z} + z\,a\,e^{-a\,t}\left(1 - e^{-a\,t}\right)^{z-1}\right] & \text{if } t \le t_1 \\ N_P\left[c_3\frac{k_0}{k_3}e^{-\frac{k_0}{k_3}t}\left(1 - e^{-a\,t}\right)^{z} + z\,a\,e^{-a\,t}\left(1 - e^{-a\,t}\right)^{z-1}\left(\frac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t}\right)\right] & \text{if } t > t_1 \end{cases} \tag{3.20}$$

Here, the integer z is the number of mutations in the cell, in which cancer necessarily starts in the cell. In this model, we assume that all daughter cells inherit all maternal mutations.

The function of age distribution of cancers for three or more mutations per cell is shown in Fig. 3.5. We used the following parameters to plot the graph: $R_0 = 100$; $k_0 = 0.1$; $k_2 = 0.2$; $k_3 = 0.4$.

The number of cancers that will happen in a group of N_P people in the time interval $[t; t + dt]$, is the product of the age distribution function $Dz(t)$ and the length of the interval dt.

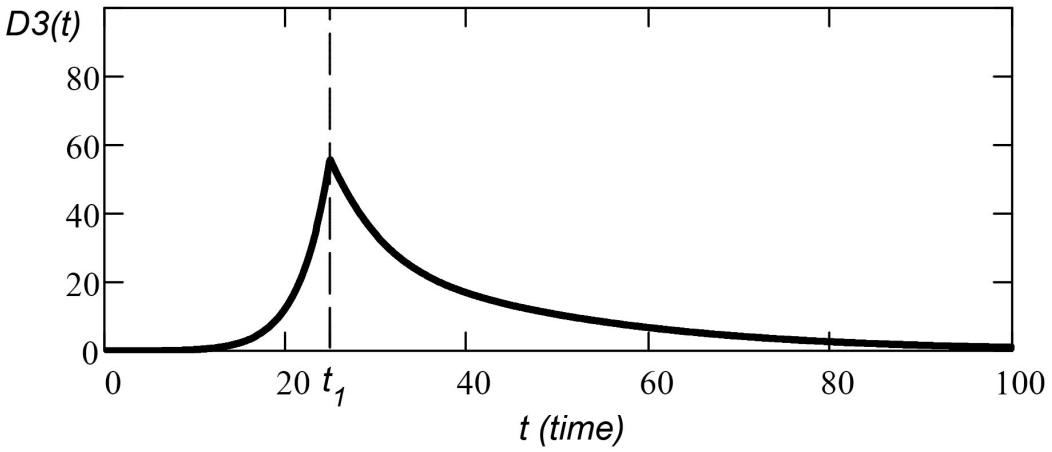

Рис. 3.5. The function of the age distribution of cancers for the cell population (provided that cancer occurs with three or more mutations in the cell).

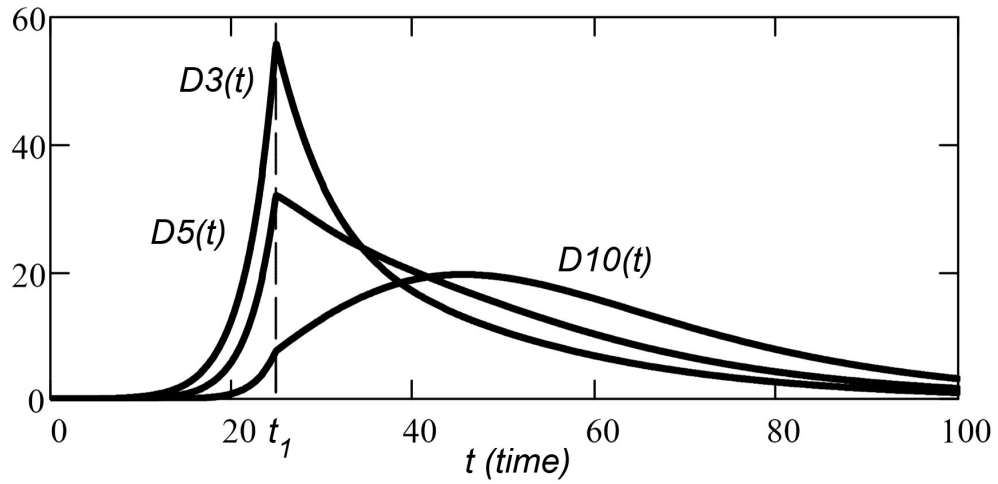

Рис. 3.6. The function of age distribution of cancers for three mutations per cell (the function $D3(t)$), for five mutations per cell (the function $D5(t)$), for ten mutations per cell (the function $D10(t)$).

The functions of the age distribution of cancers for different numbers of mutations are shown in Fig. 3.6 and 3.7. On the graphs we see a characteristic kink point of the function at the time $t = t_1$, which is smoothed only for distributions with a large number of mutations.

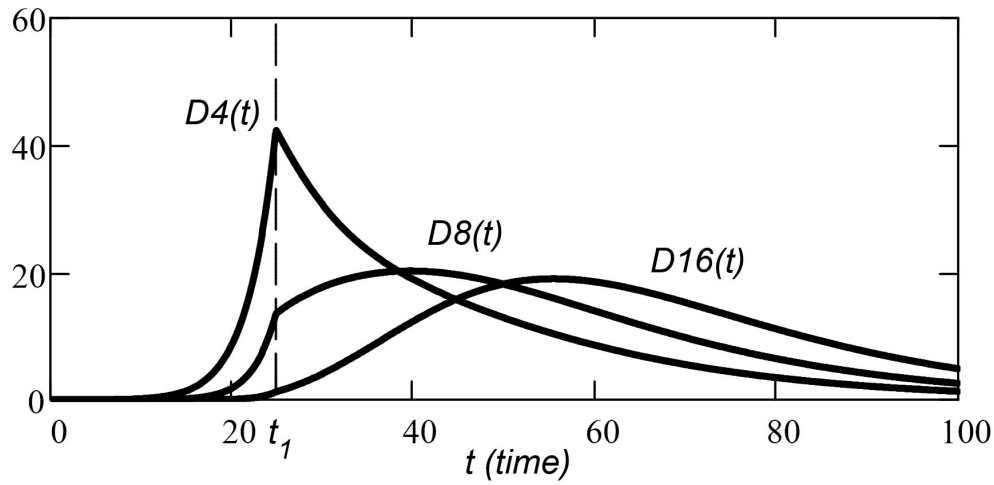

Рис. 3.7. The function of age distribution of cancers for four mutations per cell (the function $D4(t)$), for eight mutations per cell (the function $D8(t)$), for sixteen mutations per cell (the function $D16(t)$).

4. Model of Delayed Carcinogenic Action

The model of delayed carcinogenic action (DCA-model) assumes that the object of carcinogenic effect is not a single cell, but the entire cellular tissue. That is, the object of influence is an organ (tissue) or (which is mathematically the same) an individual person. In addition, the model has an age shift – the carcinogenic effect does not begin at age zero, but after some time, when a person reaches a certain age.

We will designate as T_S the time when the carcinogenic effect starts. For the occurrence of cancer, one initiating event is enough (what kind of event it is, we do not know. The average number of carcinogenic events that happen to the average person per unit of time is a. This is a very small value, and the likelihood of getting sick is small, so people get sick at different times. Until the age of T_S the number of cancers in the group is zero. After the time T_S the number of cancers begins to increase gradually.

The incidence growth graph looks like a function of $N1(t)$:

$$N1\,(t) = \begin{cases} 0 & \text{if } t \le T_S \\ N_0\left(1 - e^{-a\left(t - T_S\right)}\right) & \text{if } t > T_S \end{cases} \qquad (4.1)$$

Here N_P is the number of people in the considered population (group). In our case (for example) $N_P = 1000$. The age distribution of cancers $D1(t)$ is the time derivative of the function $N1(t)$:

$$D1\,(t) = \frac{d\left(N1(t)\right)}{dt}$$

$$D1(t) = \begin{cases} 0 & \text{if } t \le T_S \\ N_P\, a\, e^{-a(t-T_S)} & \text{if } t > T_S \end{cases} \qquad (4.2)$$

The graph of the function $D1(t)$ is shown in Fig. 4.1. For plotting we used the following parameters: $a = 0.05; N_P = 1000; T_S = 30$.

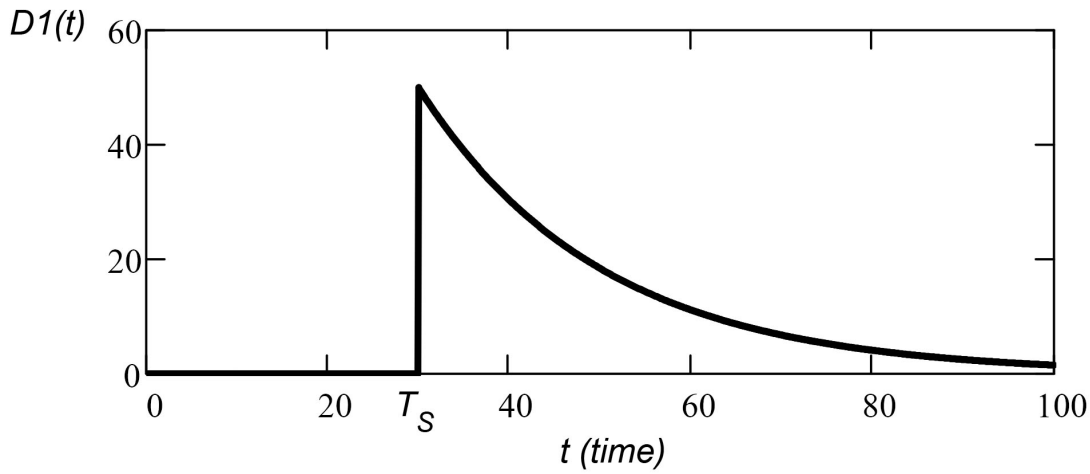

Рис. 4.1. Age distribution of cancers for a population of 1000 people (with the time shift $T_S = 30$).

In this distribution, the age of T_S (age of onset carcinogenic effect) is the same for all people. This is not true in real life. Some people have a slightly longer life expectancy than the average life expectancy, while others have a slightly shorter life expectancy. Therefore, the age of onset of carcinogenic effects is also different for everyone. In practice, the sharp peak of the distribution should be somewhat blurred along the time axis.

We assume that the time of the onset of the carcinogenic effect has a normal (Gaussian) distribution $G1(t)$ (with the standard deviation parameter σ) over the entire group of people:

$$G1(t) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(t-T_S)^2}{2\sigma^2}} \qquad (4.3)$$

To find the age distribution of cancers $D2(t)$ with a individual time shift for each person in the population, it is necessary to sum the distributions of $D1(t)$ for all people, taking into account the distribution of people with different ages $T_{S\,i}$.

Let's split the time interval $[0; \infty]$ into small time intervals dt. Then:

$$D2(0) = D0(0)\, G1(0),$$

where $D0(t)$ is the function $D1(t)$ without time shift:

$$D0(t) = N_P\, a\, e^{-a\,t}$$

At the moment o $t = dt$:

$$D2(dt) = D0(0)\,G1(dt) + D0(dt)\,G1(0)$$

At the moment o $t = 2dt$:

$$D2(2dt) = D0(0)\,G1(2dt) + D0(dt)\,G1(dt) + D0(2dt)\,G1(0)$$

etc.

When dt tends to zero, we get:

$$D2(T) = N_P \frac{a\,e^{-a\,T}}{\sigma\sqrt{2\pi}} \int_0^T e^{a\,t}\left(e^{-\frac{(t-T_S)^2}{2\sigma^2}}\right) dt \qquad (4.4)$$

For this distribution (as well as for all distributions considered earlier), the normalization condition must be satisfied:

$$\int_0^\infty D2(t)\,dt = N_P \qquad (4.5)$$

The graph of the function $D2(t)$ (age distribution of cancers in a population of 1000 people) is shown in Fig. 4.2.

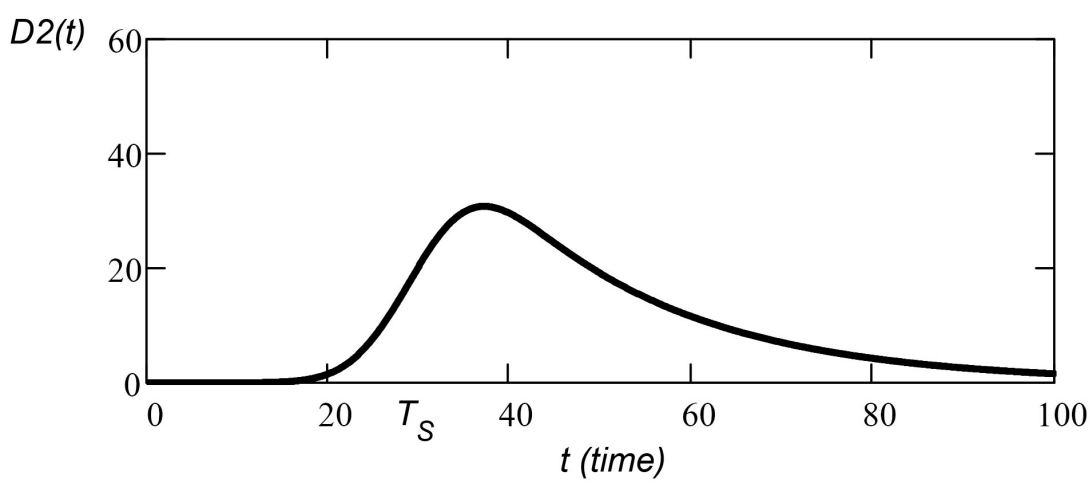

Рис. 4.2. Function $D2(t)$ (age distribution of cancers in a population of 1000 people). Age shift is $Ts = 30$.

When plotting the graph, we used the following parameters: $a = 0.05$; $N_P = 1000$; $T_S = 30$; $\sigma = 13.4$. Compare the graph in Fig. 4.2 with the graph in Fig. 4.1.

Part II
Statistical analysis of age distributions of cancers

5. Practical datasets

The data presented in [6] were selected as datasets for statistical analysis.

These are generalized data from the register of carcinogenic diseases in Belarus. Data are presented as the age distribution of the incidence of lung, stomach, colon and breast (mammary) cancer in women and the age distribution of lung, stomach, colon and prostate cancer in men.

Age distribution data are given per 100,000 people, which is convenient for numerical approximation.

All datasets, except for data on the incidence of mammary cancer, are averaged for the period from 2006 to 2010. Mammary cancer is presented for 2010, without averaging statistics.

The datasets are presented in Fig. 6.1 (and also in Figures 7.1 and 8.1) as separate black square dots. Smooth curves on the graphs are approximating functions obtained separately for each model of age distribution (the parameters of the approximating functions are described in detail below in the corresponding sections).

The graphs in the left column of Fig. 6.1 show data for female cancers, in the right column for male cancers.

6. Statistical analysis based on a simple mutation model

A simple mutational model does not correspond much to reality, but (when optimizing the approximation parameters) it demonstrates the features of the behaviour of the approximating function, which are also characteristic of a more complex mutational model.

As an approximating function, we use the age distribution function $Dz(t)$ (See. Eq. 2.37):

$$Dz(t) = N_0\, e^{-a(t-T_S)}\, a\, \frac{(a(t-T_S))^{(z-1)}}{(z-1)!}\,, \tag{6.1}$$

where z is the number of mutations in the cell, after which cancer necessarily begins.

Since cancer diagnosis is usually delayed in relation to the onset of the disease (cancer develops slowly, at the initial stage the cancer is practically invisible and is usually diagnosed after a few years), the formula includes the T_S lag time.

The function is optimized for four parameters: z, a, N_0, and T_S. The least squares method is used to select a set of parameters for which the error function $Er(z, a, N_0, T_S)$ has a minimum value:

$$Er(z, a, N_0, T_S) = \frac{1}{n}\sqrt{\sum_{k=0}^{n}\Big(D_k - Dz(t_k)\Big)^2} \tag{6.2}$$

Here t_k is the age from the real dataset; D_k is the value of the age distribution from the real dataset for the age of t_k; $Dz(t_k)$ is the theoretical value of the age distribution for the age t_k.

The optimization of the function Er is carried out programmatically, by the Levenberg-Marquardt method [7] [8] for each fixed value of z (since the z parameter is a positive integer).

To compare the magnitude of the error for different age distributions, for the obtained (optimized) parameters, we calculate the value of the relative error $Er_{\%}$ (as a percentage of the maximum value D_{max} from the real dataset):

$$Er_{\%}(z, a, N_0, T_S) = \frac{100}{n}\sqrt{\sum_{k=0}^{n}\left(\frac{D_k - Dz(t_k)}{D_{max}}\right)^2} \tag{6.3}$$

Here D_{max} is the maximum value of the age distribution function from the real dataset.

The fixed maximum value is used due to the fact that in the real and theoretical distribution the points of the curves (points that are not close to the point of mathematical expectation) are very close to zero. This leads to a strong scatter of the relative error and the impossibility of comparing the optimization quality for different age distributions.

The obtained optimal parameters z, a, N_0 and T_S and the values of the absolute and relative errors (the functions Er and $Er_{\%}$) are shown in Table 6.1 (separately for female and male cancers).

In the bottom line of the table (for women and for men) the parameters averaged over all types of cancer are shown in bold.

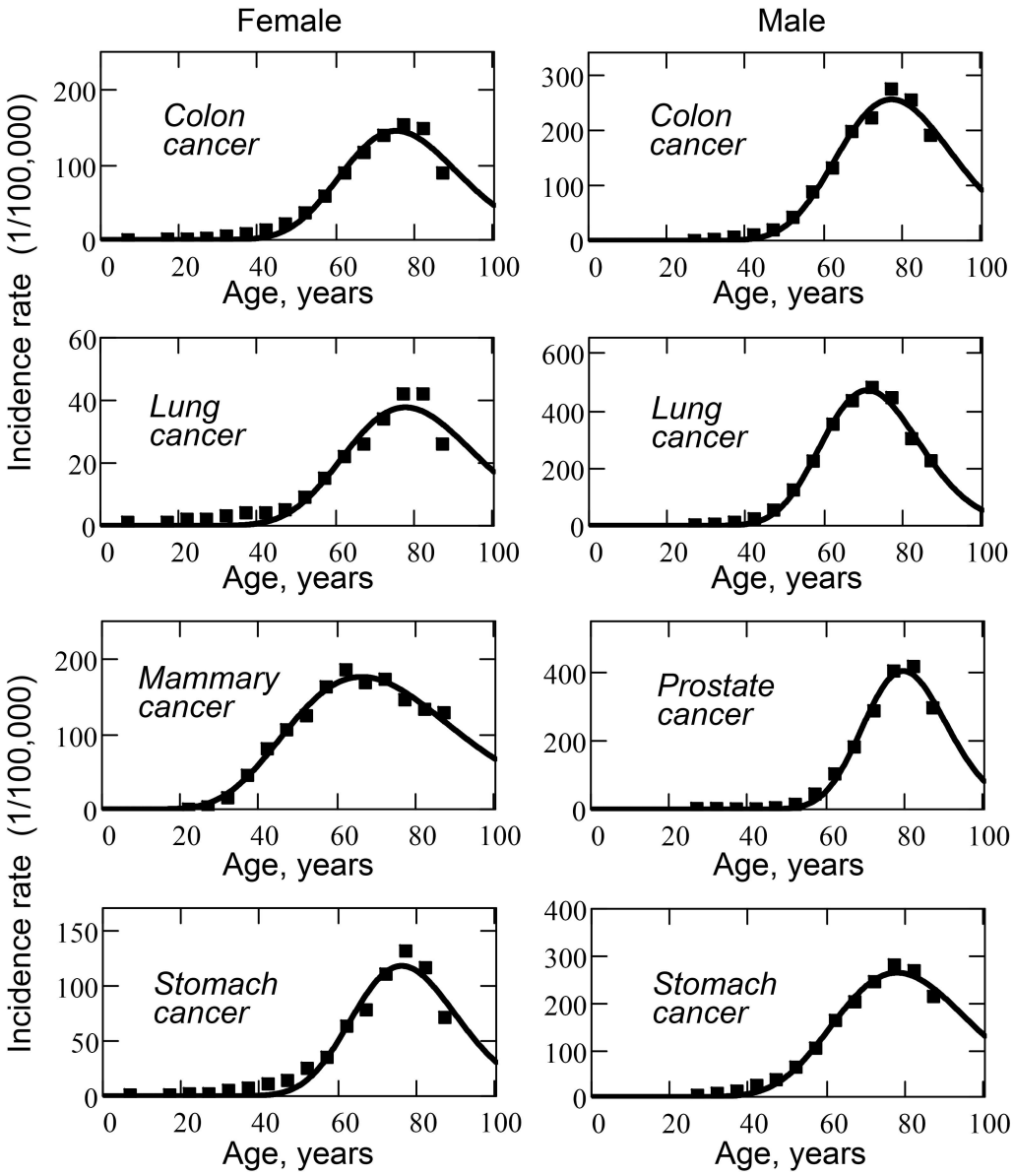

Рис. 6.1. Age distributions of cancers per 100,000 people in women (left column) and men (right column), which are approximated by the function $Dz(t)$ within the framework of a simple mutational model.

The average number of mutations required to form a cancer is 24 mutations per cell for female cancers and 35 mutations per cell for male cancers. The average number of mutations per unit time is 0.30 mutations per year per cell in women and 0.45 mutations per year per cell in men.

The average (for all types of cancer) number of cells in the group forming a given age distribution of cancers is 5,075 cells for women and 11,556 cells for men.

Table 6.1.
Fitting parameters for a simple mutation model.

Female	z	T_S	a	N	Er	$Er_{\%}$
Colon	28	0.76	0.34	5,422	2.07	1.35
Lung	23	0.25	0.29	1,566	0.78	1.86
Mammary	10	2.88	0.14	9,367	1.95	1.05
Stomach	33	0.53	0.43	3,945	2.06	1.57
Averaged	**24**	**1.11**	**0.30**	**5,075**	**1.72**	**1.46**

Male	z	T_S	a	N	Er	$Er_{\%}$
Colon	28	1.10	0.36	9,384	2.78	1.02
Lung	33	0.49	0.46	14,640	3.57	0.75
Prostate	56	0.34	0.70	10,820	3.95	0.94
Stomach	21	1.13	0.26	11,380	2.71	0.97
Averaged	**35**	**0.77**	**0.45**	**11,556**	**3.25**	**0.92**

The approximating functions (the solid curved lines in graphs) corresponding to the optimal parameters of the function Er are shown in Fig. 6.1. The dark points in the graphs are datasets of the real age distribution of cancers.

7. Statistical analysis based on a complex mutation model

A complex mutation model takes into account the growth of the cell population and the transfer of mutations obtained by the parent cells to the daughter cells.

Population growth occurs in two stages: at the initial stage the population size grows exponentially, at the second stage the population has limited growth and the population size is gradually approaching a certain limit.

At the initial moment of time ($t = 0$) the population consists of one cell. Then the population increases. The final population size is given by the expression R_0/k_0. The R_0 parameter is the value of the available energy (food) resource (which is available to the population per unit of time). The R_0 parameter defines the final number of cells in the population at the time $t = \quad infty$.

We can estimate the number of N_G cells in the epithelial cell tissue of an adult organ if we know the anatomical size of the human organ and the size of an individual cell [9].

In this work, we used the following estimates:
Colon: $N_G = 2.0 \cdot 10^8$
Lung: $N_G = 2.0 \cdot 10^{10}$
Stomach: $N_G = 2.4 \cdot 10^7$
Mammary: $N_G = 5.0 \cdot 10^7$
Prostate: $N_G = 5.6 \cdot 10^5$

Parameter R_0 is a constant environmental parameter. The parameters k_0, k_1, k_3 are biological parameters of the cell population (see Section 2). These parameters are fixed and do not change during calculations.

The coefficients k_0, k_1, k_3 are chosen so that the growth curve corresponds to our ideas about the growth rate of the human body (the maximum growth rate is reached at the age of 10-15 years). In our case, the following values of the coefficients are set: $k_0 = 0.25$, $k_1 = 0.5$, $k_3 = 0.2$. The R_0 parameter is set so that the R_0/k_0 ratio would set the required number of cells in the cell population of the adult organ tissue.

For a complex mutational model, we use the function $Dz(t)$ as the function of age distribution of cancers (see Eq. 3.20):

$$Dz(t) = \begin{cases} N_P\, e^{b(t-T_S)} \Big[b \left(1 - e^{-a(t-T_S)}\right)^z + \\ \qquad + z\, a\, e^{-a(t-T_S)} \left(1 - e^{-a(t-T_S)}\right)^{z-1} \Big] & \text{if } t \le t_1 \\ N_P \Big[c_3 \frac{k_0}{k_3} e^{-\frac{k_0}{k_3}(t-T_S)} \left(1 - e^{-a(t-T_S)}\right)^z + \\ \qquad + z\, a\, e^{-a(t-T_S)} \left(1 - e^{-a(t-T_S)}\right)^{z-1} \left(\frac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}(t-T_S)} \right) \Big] & \text{if } t > t_1 \end{cases} \tag{7.1}$$

The time t_1 and the constant of integration c_3 are determined by the initial parameters of the population development: R_0, k_0, k_1 и k_3. As before, in the simple mutation model, the delay parameter T_S is also optimized – we believe that cancer is diagnosed not at the time of the onset of the disease, but after a few years (after the lapse of time T_S since the onset of the disease).

We optimize the function by four parameters: z, a, N_P и T_S.

The least squares method is used to select a set of parameters for which the error function $Er(z, a, N_P, T_S)$ has a minimum value:

$$Er(z, a, N_P, T_S) = \frac{1}{n} \sqrt{\sum_{k=0}^{n} \left(D_k - Dz(t_k)\right)^2}$$

Here t_k is the age from the real dataset; D_k is the value of the age distribution from the real dataset for the age of t_k; $Dz(t_k)$ is the theoretical value of the age distribution for the age t_k.

To compare the magnitude of the error for different age distributions, for the obtained (optimized) parameters, we calculate the value of the relative error $Er_{\%}$ (as a percentage of the maximum value D_{max} from the real dataset):

$$Er_{\%}(z, a, N_P, T_S) = \frac{100}{n}\sqrt{\sum_{k=0}^{n}\left(\frac{D_k - Dz(t_k)}{D_{max}}\right)^2} \qquad (7.2)$$

As before (for a simple mutation model), the optimization of the Er function is carried out programmatically, using the Levenberg-Marquardt method for each fixed value of z (because the z parameter is a positive integer).

Table 7.1.
Approximation parameters for a complex mutation model.

Female	z	T_S	a	N_P	Er	$Er_{\%}$
Colon	82	0.16	0.06	$3.2 \cdot 10^{-5}$	2.53	1.65
Lung	53	0.34	0.05	$1.5 \cdot 10^{-7}$	0.86	2.05
Mammary	17	0.94	0.04	$2.1 \cdot 10^{-4}$	1.87	1.01
Stomach	135	0.02	0.06	$2.0 \cdot 10^{-4}$	1.89	2.48
Averaged	**72**	**0.37**	**0.05**	$\mathbf{6.6 \cdot 10^{-5}}$	**1.79**	**1.78**

Male	z	T_S	a	N_P	Er	$Er_{\%}$
Colon	96	0.08	0.06	$5.7 \cdot 10^{-5}$	3.57	1.30
Lung	203	0.02	0.08	$8.1 \cdot 10^{-7}$	6.40	1.34
Prostate	536	0.01	0.08	$2.4 \cdot 10^{-2}$	5.80	1.39
Stomach	50	0.17	0.05	$5.8 \cdot 10^{-4}$	3.46	1.23
Averaged	**221**	**0.07**	**0.07**	$\mathbf{6.2 \cdot 10^{-3}}$	**4.81**	**1.32**

The obtained optimal parameters z, a, N_P and T_S, and the absolute and relative errors (the functions Er and $Er_{\%}$) are shown in Table 7.1 (separately for female and male cancers).

In the bottom line of the table 7.1 (for women and for men) the parameters averaged over all types of cancer are shown in bold.

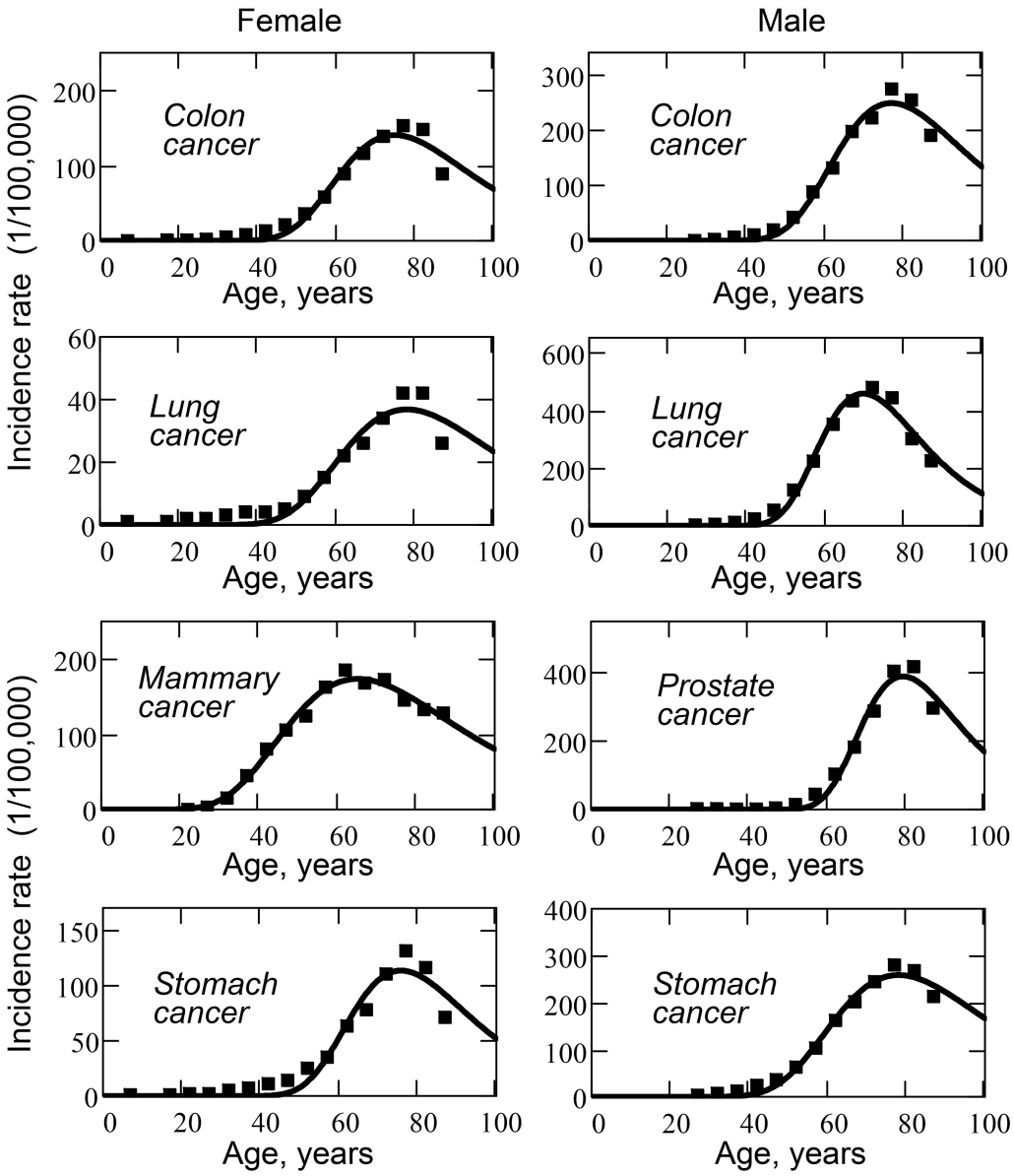

Рис. 7.1. Age distributions of cancers per 100,000 people in women (left column) and men (right column), which are approximated by the function $Dz(t)$ within the framework of a complex mutational model.

The average number of mutations required to form a cancer is 72 mutations per cell for female cancers and 221 mutations per cell for male cancers. The average number of mutations per unit of time is 0.05 mutations per year per cell for women and 0.07 mutations per year per cell for men.

The approximating functions (the solid curved lines in graphs) corresponding to the optimal parameters of the function Er are shown in Fig. 7.1. The dark points in the graphs are datasets of the real age distribution of cancers.

8. Statistical analysis based on model of delayed carcinogenic action

In the model of delayed carcinogenic action (a single event with a time shift), to approximate the age distribution of cancers, we use the function $D2(t)$ (see. Eq. 4.4):

$$D2(T) = N_P \frac{a\, e^{-a\,T}}{\sigma_1 \sqrt{2\pi}} \int_0^T e^{a\,t} \left(e^{-\frac{(t-T_S)^2}{2\sigma_1{}^2}} \right) dt \qquad (8.1)$$

Here t and T are the same variables. We use different notation to avoid confusion with the limits of integration. The error function is minimized by choosing the parameters: a, N_P, T_S, σ:

$$Er(,a, N_P, T_S, \sigma) = \frac{1}{n} \sqrt{\sum_{k=0}^{n} \left(D_k - D2(t_k) \right)^2}$$

In this case, the T_S time is the sum of the average age of onset of carcinogenic effect and the time from the onset of cancer to diagnosis of the disease.

As before, we evaluate the function of relative error $Er_{\%}$ (as a percentage of the maximum value of D_{max} from the dataset of the real age distribution).

$$Er_{\%}(a, N_P, T_S, \sigma) = \frac{100}{n} \sqrt{\sum_{k=0}^{n} \left(\frac{D_k - D2(t_k)}{D_{max}} \right)^2} \qquad (8.2)$$

The obtained optimal parameters a, N_P and T_S and the absolute and relative error values (the functions Er and $Er_{\%}$) are shown in Table 8.1 separately for female and male cancers.

In the bottom line of the table (for women and men), the parameters averaged over all types of cancer are shown in bold.

In this model, the carcinogenic effect acts not on the cell, but on an individual person (anatomical organ). The average (for all types of cancer) number of people in the group forming a given age distribution of cancers is 6,935 people for women and 11,550 people for men. The average number

of carcinogenic events (parameter a) per person per year is 0.11 events for women and 0.11 events for men. The average time T_S (the sum of the average age of onset of carcinogenic effect and the time from the onset of cancer to its diagnosis) is 64 years for women and 69 years for men.

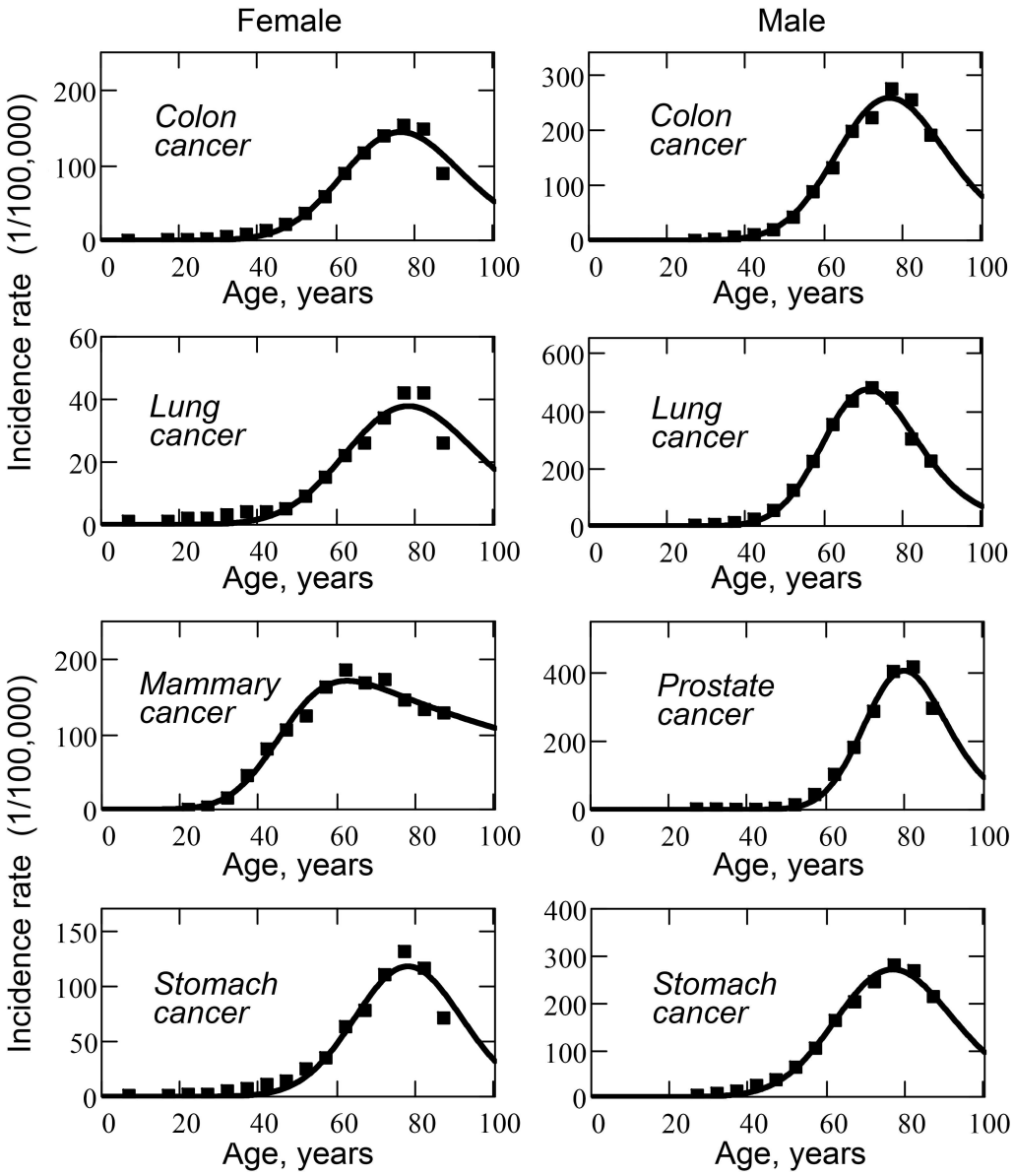

Рис. 8.1. Age distributions of cancers per 100,000 people in women (left column) and men (right column), which are approximated by the function $D2(t)$ within the framework of model of delayed carcinogenic action (see Eq. 4.4).

Table 8.1.
Approximation parameters for the model of delayed carcinogenic action.

Female	N_P	T_S	a	σ	Er	$Er_{\%}$
Colon	5,830	67	0.09	13	1.95	1.27
Lung	1,590	69	0.09	14	0.74	1.75
Mammary	16,310	46	0.01	10	2.22	1.20
Stomach	4,010	74	0.24	13	1.98	1.51
Averaged	**6,935**	**64**	**0.11**	**13**	**1.72**	**1.43**

Male	N_P	T_S	a	σ	Er	$Er_{\%}$
Colon	9,400	69	0.10	12	2.81	1.03
Lung	15,100	64	0.10	10	3.13	0.65
Prostate	11,200	74	0.13	9	3.65	0.87
Stomach	10,500	69	0.10	13	2.36	0.84
Averaged	**11,550**	**69**	**0.11**	**11**	**2.99**	**0.85**

The approximating functions (the solid curved lines in graphs) corresponding to the optimal parameters of the function Er are shown in Fig. 8.1. The dark points in the graphs are datasets of the real age distribution of cancers.

Part III
Discussion

9. Mutational models

The first thing that catches your eye when examining table 7.1 is the low number of people in a group (N_P parameter), which is necessary to obtain a real age distribution.

The size of the group should not be lower than the maximum value in the real dataset, and in our table, the size of the group is less than one person.

Maybe there is a very large error in our estimate of the number of cells in epithelial tissue?

In formula 7.1, we can put the product $N_P N_G$ outside the brackets (this is the product of the number of people in a group and the number of cells in the cellular tissue of an adult organ). This means that when approximating, the product $N_P N_G = const$. Therefore, if we increase (decrease) one value (for example, N_P) tenfold, then the other value (N_G) will decrease (increase) tenfold.

Even if we reduce the number of cells in tissue by ten thousand times (suppose we were wrong in our estimates by four orders of magnitude), we will not be able to bring the number of people in a group to a logically acceptable value (to the value of the maximum number from real age distribution).

Let's show this with an example. For prostate cancer, to increase the size of the group to 2,400 people, we need to reduce the number of cells involved in the formation of cancer to 6 cells (rounded off). This means that we have lowered our estimate a hundred thousand times – see the numbers at the beginning of Section 7.

The difference between the size of the group and the number for which the real age distribution is compiled (this is 100,000 people) can be explained with reference to a hereditary predisposition to oncological diseases (about 2,400 out of 100,000 people have a genetic predisposition to cancer).

It is more difficult to explain the small number of cells in the tissue that are susceptible to mutations. It turns out that mutations occurred (and were inherited) in only six cells of the adult organ. The rest of the cells did not feel at all the damaging effect that causes mutations. These six cells are daughter cells of one or more parent cells. This means that the damaging event, that made these parental cells susceptible to mutations, did not occur at the beginning of organ development (in childhood), but closer to old age (that is, there is a time delay in the onset of the damaging effect). This contradicts the mutational hypothesis, which states that such events should be approximately evenly distributed over time (some cells inevitably have to be damaged during

the childhood of the patient).

The second argument against mutational hypothesis is the absence of a delay (T_S parameter) in the diagnosis of the disease (in the considered models). As medical practice shows, some time elapses between the onset of cancer and its diagnosis (in our complex mutational model this time is T_S). Physicians estimate the interval to be several years. In our case (for a complex mutational model), there is no such delay. When optimizing the parameters, the time interval T_S is constantly decreasing, and the error function has a minimum when T_S tends to zero. This happens only because we have introduced a logical constraint that the lag is a positive number. The table value T_S is not zero due to the fact that the number of mutations z is an integer, and, as an optimum, the table contains the maximum z for which T_S is not negative yet. If we allow negative values of T_S, then for the error function we can get better values, and this also indicates the logical inconsistency of the mathematical mutational model.

The absence of delay (minimum of the error function for a non-zero value of T_S) is characteristic of both simple and complex mutational model.

There are articles in which, based on real age distributions of cancers, the number of cell mutations required for cancer formation is estimated. For example, in [10], the Weibull distribution is used as a mathematical model, but the equations of the growth of the cell population are not taken into account.

In [11], real age distributions are approximated using 16 different probability distributions (including the Weibull distribution) – to determine which of these distributions gives the best approximation accuracy. The best of those considered is the Erlang distribution mentioned in Section 2.2.

But we know that the Erlang distribution appears when the number of cells in the population remains constant – the population does not grow. This means that before the damaging effect begins to affect the cells, the cell population must grow to a certain number of cells. That is, the onset of the damaging effect has a large time lag. This again contradicts the mutation hypothesis.

Moreover, in our models, the Erlang distribution (for the simple mutation model), on average, also has a better approximation accuracy than a complex mutational model (that takes into account population growth). It turns out that the closer our mutational model is to reality, the worse it describes the real age distribution of cancer.

This is the third strong argument against mutational cancer models..

The fourth objection to the mutational models is the large value of the relative mean square error ($Er_{\%}$ parameter) compared with an alternative model of delayed carcinogenic action (compare the $Er_{\%}$ parameters in tables 7.1 and 8.1).

10. Model of Delayed Carcinogenic Action

The DCA-model was developed as an alternative to the mutational models, which cannot logically consistently explain the real age distributions of cancers.

As we said before, the model of a carcinogenic event with an age shift (with the same number of parameters of the approximating function) gives a smaller value of the relative root-mean-square error ($Er_{\%}$ parameter) compared with similar parameters for mutational models (compare the $Er_{\%}$ parameters in tables 7.1 and 8.1).

In this model the object of carcinogenic effect is not a single cell (as in mutational models), but an entire organ (cellular tissue). This allows quantitatively and without logical contradictions to assess the main parameters of the mathematical model, comparing them with the real age distribution of cancers.

For example, for the prostate cancer discussed above, the size of the group of people corresponding to the real age distribution is 11,550 people (men). This means that only this number of men out of 100,000 will get prostate cancer in their life,and this may be explained by reference to a hereditary predisposition to the disease.

The disadvantage of the model is the impossibility to separate (from the sum of two values) the delay time of cancer diagnosis (the time between the onset of cancer and the diagnosis of the disease). The time T_S in the model is the sum of the time of the onset of carcinogenic effect and time between the onset of cancer and the diagnosis of the disease.

The model does not tell us anything about the cause of cancer, but, nevertheless, it allows us to estimate some parameters of the carcinogenic damaging effect. In particular, we can estimate the average number of carcinogenic events per unit of time per anatomical organ (tissue). These data were presented in section 8.

Saint Petersburg
2021

References

[1] **Loeb L.A., Loeb K.R., Anderson J.P.**
Multiple mutations and cancer PNAS, Vol. 100, No. 3, pp. 776-781, February 2003. https://doi.org/10.1073/pnas.0334858100

[2] **Knudson A.G., JR.**
Mutation and Cancer: Statistical Study of Retinoblastoma. Proc. Nat. Acad. Sci.USA, Vol. 68, No., pp. 820-823, April 1971. https://doi.org/10.1073/pnas.68.4.820

[3] **Tetearing A.N.**
Numerical Model of Fishery Impact on Population Levels of Landlocked Salmon in Lake Onego (Russia) 2020, https://dx.doi.org/10.2139/ssrn.3443512

[4] **Erlang A.K.**
Telephone waiting times. Matematisk Tidsskrift. B, Vol. 31, p.25, Copenhagen, 1920

[5] **Tetearing A.N.**
Theory of populations. SSO Foundation, Moscow, 2012.

[6] **Antipova S.I., Antipov V.V., Shebeko N.G.**
Gender problems of oncology in Belarus. Medical News Journal, No. 3., Minsk, 2013 (in Russian).

[7] **Levenberg K.**
A Method for the Solution of Certain Problems in Last Squares. Quart. Appl. Math., Vol. 2. P. 164—168, 1944.

[8] **Marquardt D.**
An Algorithm for Least-Squares Estimation of Nonlinear Parameters. SIAM Journal on Applied Mathematics, Vol. 11 (2), pp. 431-441, 1963.

[9] **Great Medical Encyclopedia.**
Edition 3, Moscow, 1974-1989 (in Russian).

[10] **Calabrese P., Tavare S., Shibata D.**
Pretumor Progression. Clonal Evolution of Human Stem Cell Populations. American Journal of Pathology, Vol.164, Issue 4, pp. 1337-1346, 2004, https://doi.org/10.1038/s41598-017-12448-7

[11] **Belikov A.V.**
The number of key carcinogenic events can be predicted from cancer incidence. Scientific Reports, 7, (12170), 2017, https://doi.org/10.1038/s41598-017-12448-7