Cancer models and statistical analysis of age distribution of cancers

Alexandr N. Tetearing

May 7, 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 6.3 (mutations per cell for women) and 7.5 (mutations per cell for men) for the complex mutational model. The average (for four types of cancer) mutation rate is estimated at $7 \cdot 10^{-4}$ mutations per year (per cell for women) and at $1.8 \cdot 10^{-3}$ mutations per year (per cell for men).

As an alternative to mutational models, a model of delayed carcinogenic event is considered.

Contents

1. Introduction **3**

I Cancer models **4**

2. Simple mutational model **4**
2.1. Mathematical model 4
2.2. Age distribution of cancers 14

3. Complex mutational model **16**
3.1. Mathematical model 16
3.2. Age distribution of cancers 22

4. Model of delayed carcinogenic event **27**

II Statistical analysis of age distributions of cancers **30**

5. Practical datasets **30**

6. Statistical analysis based on a simple mutational model **30**

7. Statistical analysis based on a complex mutational model **33**

8. Statistical analysis based on model of delayed carcinogenic event **37**

III Discussion **40**

9. Mutational models **40**

10. Model of delayed carcinogenic event **42**

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 and 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.

the question of the minimum required number of cellular mutations leading to malignant transformation 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.

Thus, we can approximately find out 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

This simple mathematical model is based on the assumption that the age distribution of cancer is independent of the growth function of the anatomical organ (tissue). This model is an approximate model, since it does not take into account the shape of the growth function of the cell population.

In a simple mutation model, we consider a group of N_P people. Each person from the group has a certain anatomical organ (tissue), which consists of separate identical cells. At the initial moment of time $t = 0$, the cell population of the organ consists of one cell. Over time, the number of cells increases in accordance with the organ growth function $N(t)$.

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 cells in a short time dt is $N\,a\,dt$.

We assume that, with a certain number of mutations accumulated by the cell (we 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, and the number of registered cancer patients in this group.

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.

Take a look at Fig. 2.1. The figure schematically shows the development of cell populations in a group of 27 people. In each individual person, the growth of an organ (cell population) occurs in accordance with the growth function $N(t)$. This function is shown as a bar chart on the left side of Figure 2.1.

We consider the state of the cell population discreetly – at regular intervals dt. In this case, $dt = 1$.

At the moment of time $t = 0$, the organ (cellular tissue) consists of one cell. An organ consists of two cells at time $t = 1$, four cells at time $t = 2$, and eight cells at time $t = 3$. This growth function has been chosen solely for example. In reality, the function may have other parameters.

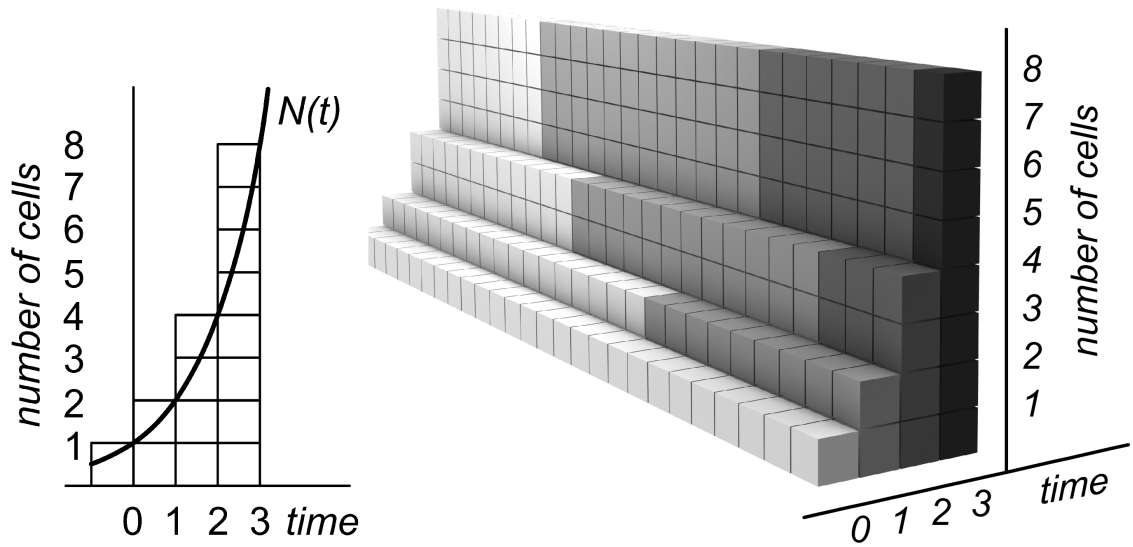

Figure 2.1. Scheme for obtaining mutations in a group of 27 people with the same functions $N(t)$ of population growth .

Thus, at the moment of time $t = 0$, in a group of 27 people there are 27 separate cells (27 people have 27 identical organs, and each organ contains one cell). These are 27 light-colored cells in the first row.

By the time $t = 1$, the number of cells has doubled, and the first mutation has occurred in some of the cells. We consider (for example) that a third of all cells undergo mutation per unit time. That is, in one cell per unit of time, on average, 0.33 mutations occur (in a group of three cells, one cell receives a mutation). 18 cells that have received a mutation at the time interval $[0; 1]$ are shown (in two rows) in light gray. In total, at the moment of time $t = 1$, the figure shows 54 cells (36 of them have no mutations).

For this model, we believe that mutations are inherited from the parent cell to the daughter cells.

By the time $t = 2$, the number of cells will double again – up to 108 (4 rows of 27 cells), and some of the cells will undergo mutation. One third of cells with one mutation will receive the second mutation (cells with two mutations are shown in dark gray). One third of cells that do not have mutations will receive the first mutation. Thus, 12 cells (four rows of 3 cells) by the time $t = 2$ will have two mutations (dark gray cells); 48 cells (4 rows of 12 cells) will each have one mutation (light gray cells).

By the time $t = 3$, the number of cells will double again – up to 216 cells (8 rows of 27 cells), and some of the cells will mutate again. A third of cells with two mutations will receive a third mutation (cells with three mutations

are shown in black). A third of cells with one mutation will receive the second mutation (cells with two mutations are shown in dark gray). One third of cells that do not have mutations will receive the first mutation. Thus, by the time $t = 3$, 8 cells (8 rows of 1 cell) will have three mutations (black cells in the figure); 48 cells (8 rows of 6 cells) will have two mutations (dark gray cells), 96 cells (8 rows of 12 cells) will each have one mutation (light gray cells); 64 cells (8 rows of 8 cells) will not have mutations (these are the lightest cells in the picture).

Each vertical column in Fig. 2.1 consists of cells of the same color (these are cells with the same number of mutations). Therefore, we can determine the number of people in whose anatomical organ there are cells with a certain number of mutations by looking only at the cells of the lower layer. This means that in this mathematical model, the function of the number of people in a group who have (for example) one mutation does not depend on the growth function of the cell population (the growth function of cell tissue).

In fact, the described scheme for obtaining mutations does not correspond to reality and is a great simplification. For example, in our model, at time $t = 1$, nine people have cells with one mutation – each of nine people has two cells with one mutation. In reality, mutations in a group are distributed more evenly, and it is more likely that 18 people will receive one mutated cell each.

A simplified scheme for obtaining mutations allows us to disregard the shape of the growth function of the cell population $N(t)$. A more complex and more realistic model will be discussed in the following sections.

Looking at the bottom layer of cells in Fig. 2.1, we can draw the functions $N1_P(t)$, $N2_P(t)$, and $N3_P(t)$. The $N1_P(t)$ function describes the number of people who have cells with one or more mutations. The $N2_P(t)$ function describes the number of people who have cells with two or more mutations.

The $N3_P(t)$ function describes the number of people who have cells with three or more mutations, etc. The P index in the function name means that the function describes the number of people, not the number of cells.

The functions $N1_P(t)$, $N2_P(t)$ and $N3_P(t)$ are shown in Fig. 2.2 in the form of bar charts. These functions correspond to Figure 2.1.

The functions shown in Fig. 2.2 are not smooth functions. They are depicted as steps. To get smoother (and more accurate) functions, the time interval dt should tend to zero (we record the state of the cell population at regular intervals dt). In Fig. 2.1 and 2.2 this interval is equal to one.

Consider the above functions for an arbitrary growth function of the cell population. The number of people in the group under consideration is N_P.

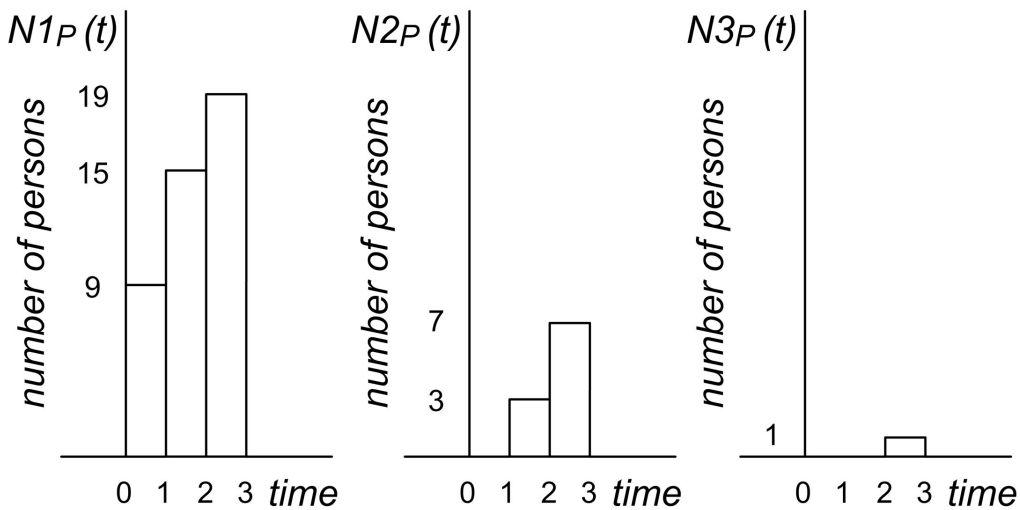

Figure 2.2. Function $N1_P(t)$ – the number of people who have cells with one or more mutations. Function $N2_P(t)$ – the number of people who have cells with two or more mutations. Function $N3_P(t)$ – the number of people who have cells with three or more mutations. The functions correspond to Fig. 2.1.

At time $t = dt$, the number of people with mutated cells is:

$$N1_P(dt) = N_P\, a\, dt \tag{2.1}$$

Here a is the average number of mutations that a cell receives per unit of time.

The number of people without mutations $N0_P$ is:

$$N0_P(dt) = N_P(1 - a\, dt) \tag{2.2}$$

At time $t = 2dt$ the number of people without mutations $N0_P$ is:

$$N0_P(2dt) = N_P(1 - a\, dt) - N_P(1 - a\, dt)\, a\, dt = N_P(1 - a\, dt)^2 \tag{2.3}$$

At time $t = 3dt$ the number of people without mutations $N0_P$ is:

$$N0_P(3dt) = N_P(1 - a\, dt)^3, \tag{2.4}$$

etc.

Thus, the number of people without mutations at time $t = T$ is the product of N_P (group size) and a definite multiplical of the negative constant a on the interval $[0; T]$:

$$N0_P(T) = N_P \prod_0^T (-a)\, dt \tag{2.5}$$

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

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

In accordance with the basic formula of multiplical calculus, the definite multiplical of the function $f(t)$ is the difference of the antiderivatives $F(t)$ 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 derivation of the formula is given in [3]. In our case, since a is a constant:

$$N0_P(T) = N_P \prod_0^T (-a)\,dt = N_P\, e^{-a\,T} \tag{2.8}$$

The number of people who have at the time t cells with one mutation (or with a large number of mutations) is:

$$N1_P(t) = N_P\,(1 - e^{-a\,t}) \tag{2.9}$$

The graph of the function $N1_P(t)$ is shown in Fig. 2.3. We used the following parameters to plot the graph: $N_P = 1000$; $a = 0.05$.

Now let's find the number $N1^*_P(t)$ of people in the group who at time $t = T$ have cells with only one mutation (no more and no less) and who do not have cells with the number of mutations exceeding one.

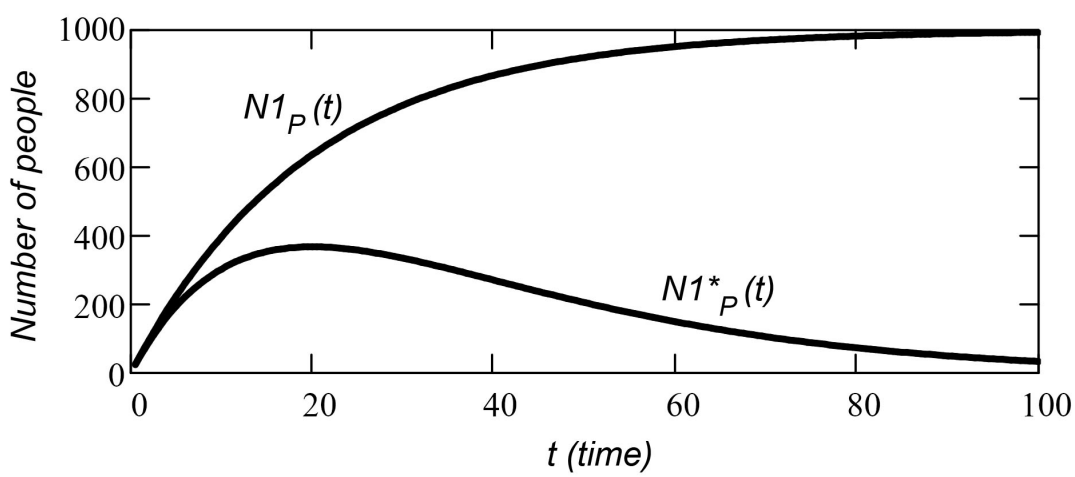

Figure 2.3. The number of people who have (at the time t) the cells with one mutation (or more) – function $N1_P(t)$ (upper curve in the figure). The number of people who have cells with only one mutation – function $N1^*_P(t)$ (bottom curve).

People who have cells with only one mutation are formed from people who did not have mutations, the number of which is described by the function $N0_P(t)$.

At each time interval dt (starting from the moment $t = dt$), there are two processes that change the number of people in the group who have cells with one mutation.On the one hand, the available number of people with one mutation decreases due to the fact that some people receive the second mutation per cell. On the other hand, the number of people who have cells with one mutation increases due to the fact that some people without mutations receive one mutation. (on each interval dt this is the increment of the function $N1_P(t)$ on the interval dt).

Since the function $N1_P(t)$ is a function that changes over time, we cannot simply use a formula in the form (2.5) to find $N1^*_P(t)$. Instead, we need to calculate the product of $dN1_P$ (the increment of the function $N1_P(t)$) and a definite multiplical on each dt interval. Let's look at Fig. 2.4.

The known function $N1_P(t)$ on the interval $[0;T]$ is bounded, positive and monotonically increasing. Divide the interval $[0;T]$ into small segments dt. The number of these segments is $n = T/dt$. In this case, $n = 4$.

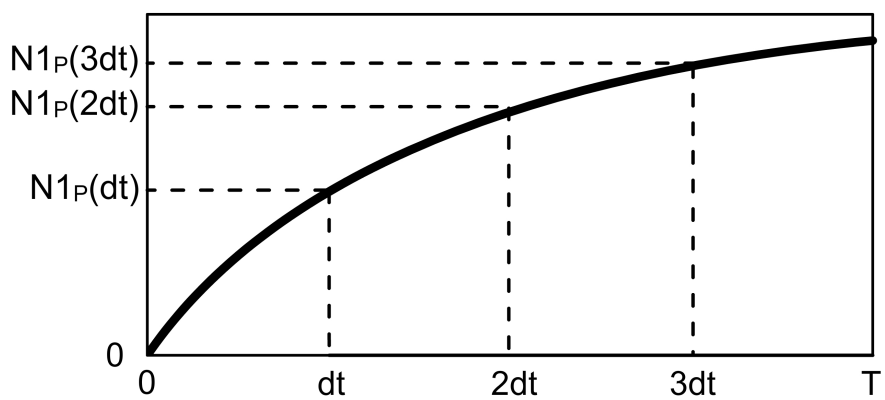

Figure 2.4. To the calculation of the function $N2_P(t)$ (the number of people in a group who have cells with two mutations).

Then the function $N1^*_P(t)$ is the sum:

$$N1^*_P(T) = N1_P(dt)\int_{dt}^{T}(-a)\,dt + \Big(N1_P(2dt) - N1_P(dt)\Big)\int_{2dt}^{T}(-a)\,dt + \\ + \Big(N1_P(3dt) - N1_P(2dt)\Big)\int_{3dt}^{T}(-a)\,dt + \Big(N1_P(4dt) - N1_P(3dt)\Big) \qquad (2.10)$$

or:

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

where $f(t)$ is the time derivative of the function $N1_P(t)$:

$$f(t) = N1_P{}'(t) = \frac{d\left(N1_P(t)\right)}{dt} \tag{2.12}$$

If dt is infinitesimal (for $n\ longrightarrow\ infty$), then it is true:

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

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

$$N1_P^*(T) = e^{-aT}\int_0^T N1_P'(t)\,e^{at}dt \tag{2.14}$$

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

$$N1_P^*(T) = e^{-aT}\int_0^T N_P\,a\,e^{-a\,t}\,e^{at}dt = e^{-aT}\,N_P\,a\int_0^T dt \tag{2.15}$$

Thus, the number of people, who have cells with only one mutation (who do not have cells with a number of mutations exceeding one), changes over time as a function of $N1_P^*(t)$:

$$N1_P^*(t) = N_P\,a\,t\,e^{-a\,t} \tag{2.16}$$

We can calculate the function $N1_P^*(t)$ (the number of people who have cells with at most one mutation) in a different way. To do this, we need to solve a differential equation:

$$dN1_P^*(t) = dN1_P(t) - N1_P^*(t)\,a\,dt \tag{2.17}$$

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

The second (negative) term on the right is the number of people that received the cells with a second mutation in the dt interval (these people are formed from people who previously had cells with only one mutation).

Dividing the equation by dt we have:

$$\frac{dN1_P^*(t)}{dt} = \frac{dN1_P(t)}{dt} - N1_P^*(t)\,a \tag{2.18}$$

We introduce the notation:

$$y = N1_P^*(t); \quad y' = \frac{dN1_P^*(t)}{dt}; \quad f' = \frac{dN1_P(t)}{dt} \tag{2.19}$$

Linear non-homogeneous differential equation:

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

is solved by the substitution $y = uv$.

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

When 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):

$$N1_P^*(t) = c_1e^{-a\,t} + e^{-a\,t}\int \frac{dN1_P(t)}{dt}\,e^{a\,t}dt \tag{2.25}$$

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

$$N1_P^*(t) = c_1e^{-a\,t} + e^{-a\,t}\int N_P\;a\,e^{-a\,t}\,e^{a\,t}dt \tag{2.26}$$

$$N1_P^*(t) = c_1e^{-a\,t} + N_P\;a\,t\,e^{-a\,t} \tag{2.27}$$

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

$$c_1 = N1_P^*(0) \tag{2.28}$$

In our case, $N1_P^*(0) = 0$. As a result, we get:

$$N1_P^*(t) = N_P\;a\,t\,e^{-a\,t} \tag{2.29}$$

So, we obtained formula (2.16) in a different way.

The number of people $N2_P$ with two mutations per cell (or more) is the difference of the functions $N1_P(t)$ and $N1^*_P(t)$:

$$N2_P(t) = N1_P(t) - N1^*_P(t) = N_P(1 - e^{-a\,t} - a\,t\,e^{-a\,t}) \qquad (2.30)$$

The graph of the function $N2_P(t)$ is shown in Fig. 2.5 (upper curve).

To find the number of people $N2^*_P$ who have cells with only two mutations (no less, no more), we will use the formula (2.14) obtained earlier. Now we need to substitute the function $N2_P(t)$ into this formula instead of the function $N1_P(t)$.

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

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

The graph of the function $N2^*_P(t)$ is shown in Fig. 2.5 (bottom curve).

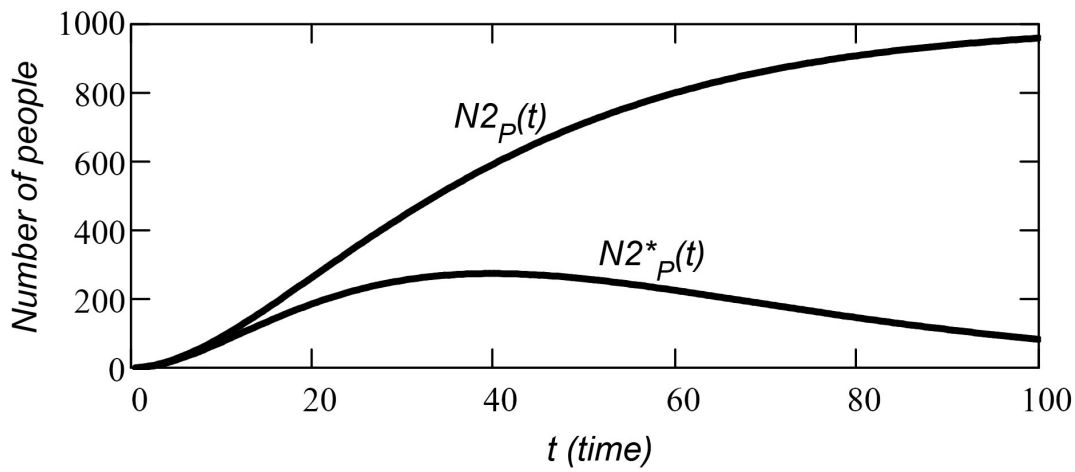

Figure 2.5. The number of people who have cells with two (or more) mutations – the function $N2_P(t)$ (upper curve in the figure). The number of people who have cells with only two mutations (who do not have cells with fewer or more mutations) – the function $N2^*_P(t)$ (bottom curve).

If we subtract the function $N2^*_P(t)$ (number of people who have cells with exactly two mutations) from the function $N2_P(t)$ (number of people who have cells with two or more mutations), we get the function $N3_P(t)$ (the number of people who have cells with three or more mutations).

$$N3_P(t) = N_P\left(1 - e^{-at} - a\,t\,e^{-at} - \frac{(a\,t)^2}{2}e^{-at}\right) \qquad (2.32)$$

To find the function $N3^*_P(t)$ (number of people who have cells with exactly three mutations), we again use formula (2.14). In this time, we need to substitute the function $N3_P(t)$ into the formula instead of the function $N1_P(t)$:

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

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

The graphs of the functions $N3_P(t)$ and $N3^*_P(t)$ are shown in Fig. 2.6.

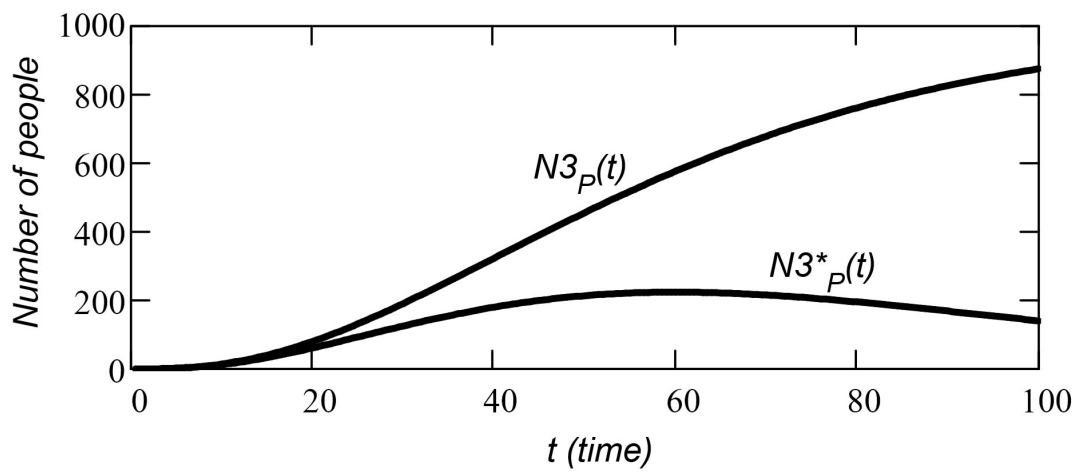

Figure 2.6. The number of people who have cells with three (or more) mutations – the function $N3_P(t)$ (upper curve in the figure). The number of people who have cells with exactly three mutations – the function $N3^*_P(t)$ (bottom curve).

Similarly, we find the functions $N4_P(t) = N3_P(t) - N3^*_P(t)$ (the number of people who have cells with four mutations or more) and $N4^*_P(t)$ (the number of people who have cells with exactly four mutations).

$$N4_P(t) = N_P\left(1 - e^{-at}\left(1 + a\,t + \frac{(a\,t)^2}{2} + \frac{(a\,t)^3}{6}\right)\right) \qquad (2.34)$$

To find the function $N4^*_P(t)$ (the number of people who have cells with exactly four mutations) we use formula (2.14). We substitute the function $N4_P(t)$ into the formula instead of the function $N1_P(t)$:

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

$$N4^*_P(t) = e^{-at}\,N_P\,\frac{(a\,t)^4}{24} \qquad (2.35)$$

The graphs of the functions $N4_P(t)$ and $N4^*_P(t)$ are shown in Fig. 2.7.

Looking at the formulas for functions $N1_P(t)$, $N2_P(t)$, $N3_P(t)$, and $N4_P(t)$, it is easy to deduce the general formula for the function $Nz_P(t)$, where z is the minimum number of mutations in a cell of an anatomical organ in one person.

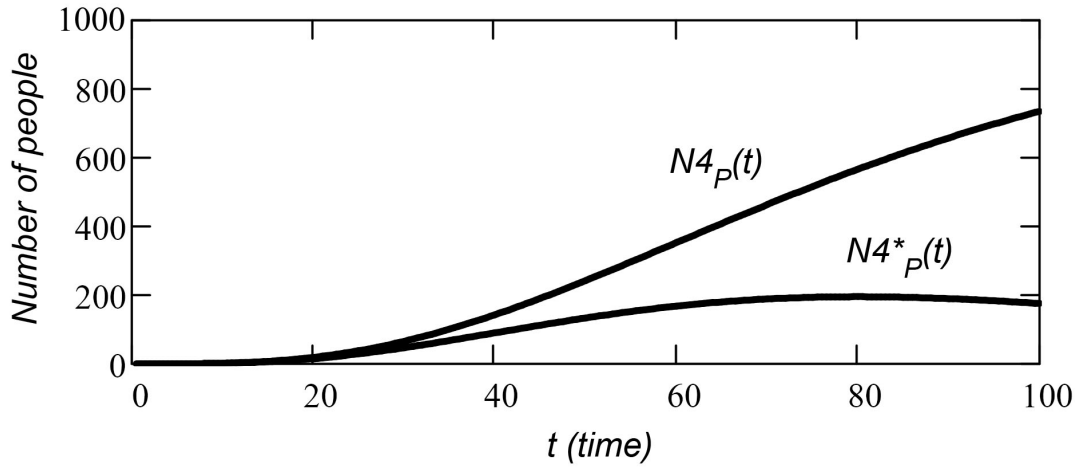

Figure 2.7. The number of people who have cells with four (or more) mutations – the function $N4_P(t)$ (upper curve in the figure). The number of people who have cells with exactly four mutations – the function $N4^*_P(t)$ (bottom curve).

For example, the designation $N8_P(t)$ corresponds to a function that describes the number of people in a group that have cells with the number of mutations from 8 or more.

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

Here N_P is the size of the group in question.

2.2. Age distribution of cancers

We agreed (this is just for example) that if a cell gets three (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 $Nz_P(t)$ (in our case, this is the function $N3_P(t)$). 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 three mutations $D3(t)$ is:

$$D3(t) = \frac{d\left(N3_P(t)\right)}{dt} = N_P\, e^{-at}\, \frac{a^3\, t^2}{2}$$

The distribution is shown in Fig. 2.8.

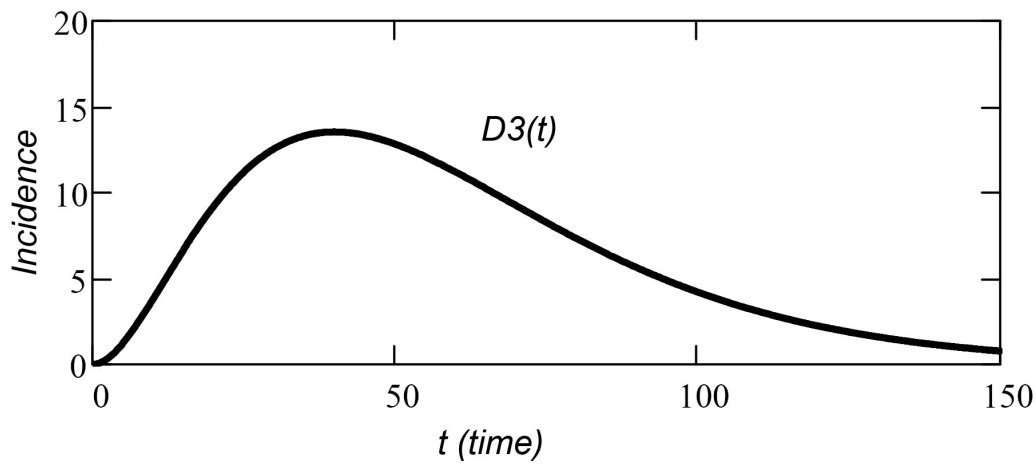

Figure 2.8. Age distribution of cancers in a group of 1000 people. The case when the disease begins with three mutations in the cell. Simple (rough) mutation model.

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:

$$Dz(t) = N_P\, e^{-at}\, a\, \frac{(a\,t)^{(z-1)}}{(z-1)!}\,, \qquad (2.36)$$

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

This is the Erlang distribution known in mathematics [4]. The Danish mathematician Agner Krarup Erlang got this distribution when he worked on the problem of determining the waiting time for a telephone exchange operator.

The graphs of the age distribution for different numbers of mutations are shown in Fig. 2.9. 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 $D8(t)$).

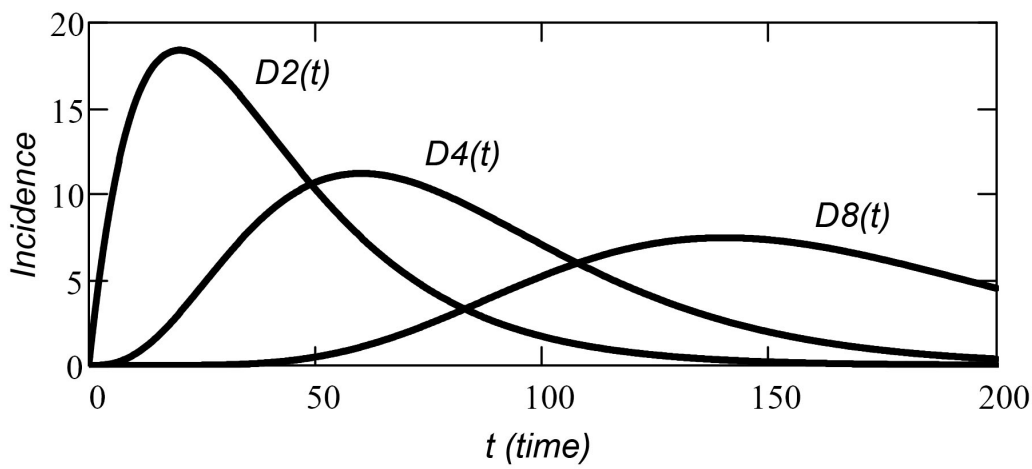

Figure 2.9. Age distribution of cancers in a group of 1000 people. The cases when the disease begins with two mutations in the cell (the function $D2(t)$), with four mutations in the cell (the function $D4(t)$), and with eight mutations in the cell (the function $D8(t)$).

For all functions considered, the number of people in a group is $N_0 = 1000$ people, the average number of mutations per cell is $a = 0.05$ mutations per unit of time.

The number of cancers in the group, in the interval $[t; t+dt]$, is the product 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, we obtained the age distribution of cancers in a group of people without taking into account the shape of the growth function of the anatomical organ (cell tissue). However, this simplest model, as we have seen, does not correspond to reality. Real cell tissue grows, increasing the number of cells by several orders of magnitude during the growth process, and the changes in growth function can affect the shape of the age distribution.

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 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).

According to the equations of population theory [5], only the ratio between the coefficients matters, but not the absolute values of the coefficients k_0, k_1, k_3. That is, the equations do not change if all three coefficients are multiplied or divided by some number. Therefore, it is often believed that the coefficient k_1 is equal to one.

The coefficient k_0 is always less than k_1, otherwise the mass of the population will decrease over time. All the equations and coefficients that characterize the cell population are described in detail in [5].

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).

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 the population) demonstrates exponential growth up to the time $t = t_1$ and limited growth after the time $t = t_1$. At the moment $t = 10$, the number of cells in the population is $N(10) = 8.6\ cdot10^4$. The population size at the time $t = t_1$ is $N(t_1) = N_R = 2.5 \cdot 10^7$. With limited growth, the number of cells gradually reaches an asymptotic plateau $N(\infty) = R0/k_0 = 5 \cdot 10^7$.

For plotting graph 3.1, we used the following parameters: $R_0 = 2.5 \cdot 10^7$; $k_0 = 0.5$; $k_1 = 1$; $k_3 = 0.44$; $t_1 = 15$.

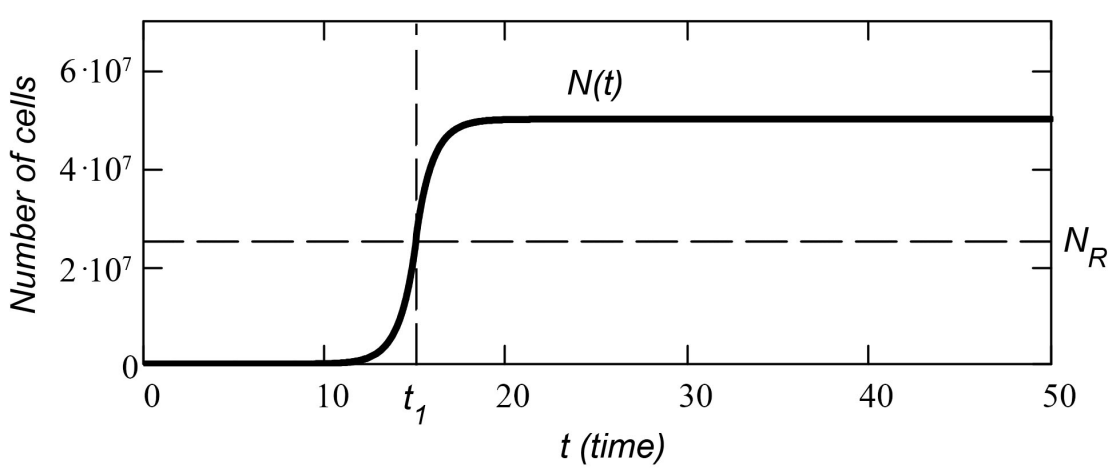

Figure 3.1. Function $N(t)$ is the number of cells in the population. The cell population has exponential growth until time $t = t_1$ and limited growth since time t_1.

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).

$$N0(T) = N(T) \prod_0^T (-a)\, dt = N(T)\, e^{-aT} \tag{3.8}$$

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 functions $N(t)$ 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:

$$N1^*(T) = N1(T) \prod_0^T (-a)\, dt \tag{3.11}$$

Here T and t is 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.

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:

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

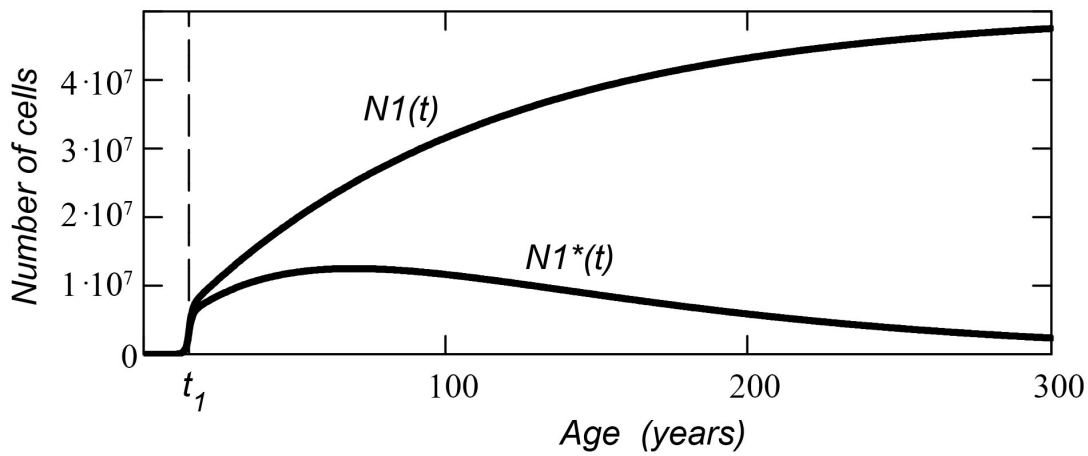

Figure 3.2. Number of cells with one or more mutations (the function $N1(t)$ and upper curve in the figure) and number of cells with exactly one mutation (the function $N1^*(t)$ and bottom curve).

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

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

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

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.

The formula for the function $N3(t)$ (the number of cells with three or more mutations) is found in a similar way.

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

Or:

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

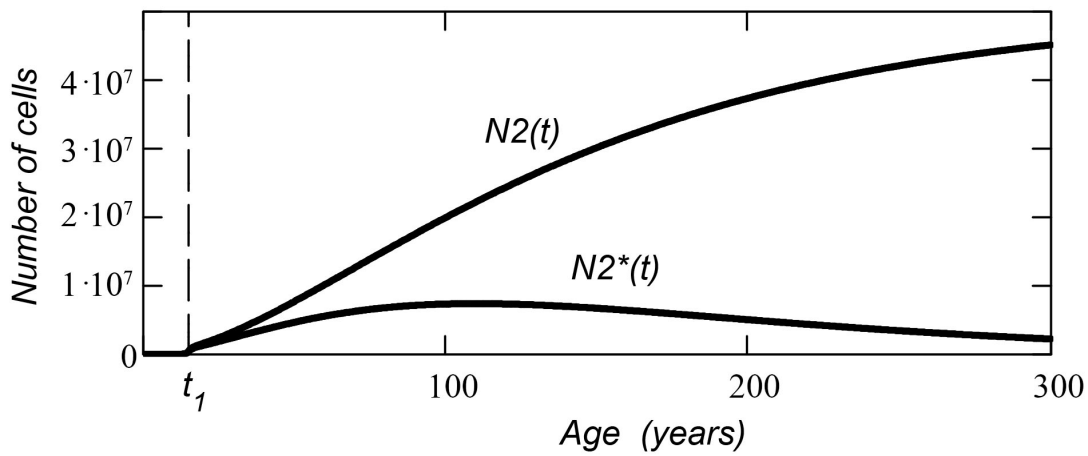

Figure 3.3. Number of cells with two or more mutations (the function $N2(t)$ and upper curve in the figure) and number of cells with exactly two mutations (the function $N2^*(t)$ and bottom curve).

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

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

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

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.

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):

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

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

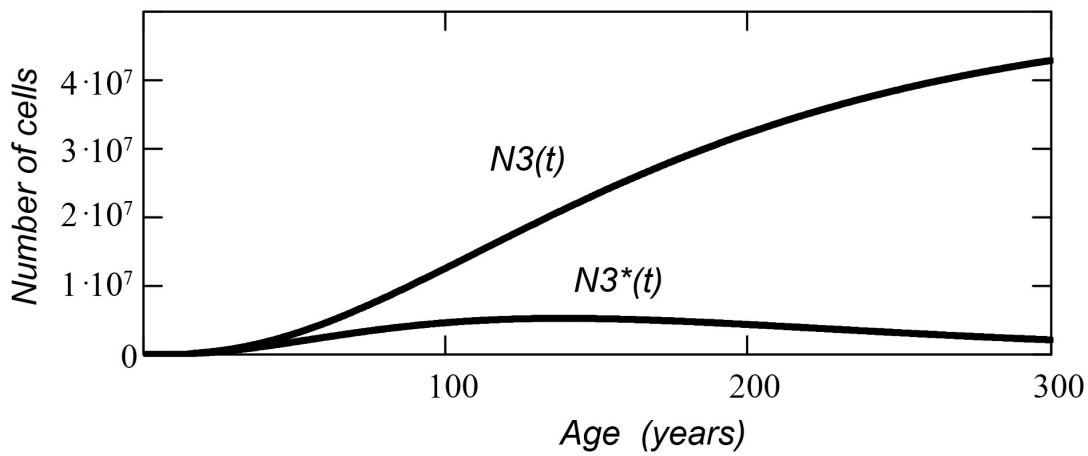

Figure 3.4. Number of cells with three or more mutations (the function $N3(t)$ and upper curve in the figure) and number of cells with exactly three mutations (the function $N3^*(t)$ and bottom curve).

3.2. Age distribution of cancers

Using the formulas in the previous section, we can get the age distribution of cancers for a group of N_P people.

We assume (for example) that cancer starts after three mutations in one cell. The number of cells with three mutations in a group of N_P people is the function $N3_P(t)$:

$$N3_P(t) = N_P N3(t) \tag{3.18}$$

The function $N3(t)$ is the number of cells with three mutations (or more) in the cell population of a separate anatomical organ (tissue).

At the initial moment of time, the number of healthy people in the group is N_P. In a short time interval dt, a certain number of cells with three mutations will appear in the population. This number is:

$$N_P\, dN3(0)\, dt\,,$$

where $N3(0)$ is the increment of the function $N3(t)$ on the first interval dt.

When the interval is small and the size of the group is large enough, the number of new cancer cells is equal to the number of diseased people. On the next dt interval, the number of healthy people is:

$$N_P\left(1 - dN3(0)\, dt\right) dN3(dt)\, dt\,,$$

We excluded from the healthy group people who fell ill in the previous time interval. At an arbitrary time moment t, the number of healthy people in the group is the product of the group size N_P and a definite multiplical

on the interval $[0; t]$ of the time derivative of the function $N3(t)$ taken with a negative sign (see formulas 2.6 and 2.7):

$$N_P \int_0^t \left(-N3'(t) \right) dt = N_P\, e^{-N3\,(t)} \tag{3.19}$$

The number of sick people $N3_P\,(t)$ is equal to:

$$N3_P\,(t) = N_P \left(1 - e^{-N3\,(t)} \right) \tag{3.20}$$

To get the age distribution of cancer $D3(t)$ in a group, we need to find the time derivative of the function $N3_P\,(t)$:

$$D3(t) = \frac{d\left(N3_P\,(t)\right)}{dt} = -N_P\, e^{-N3\,(t)}\; N3'(t) \tag{3.21}$$

The age distribution $Dz(t)$ of cancers in a group of N_P – for the case when the number of mutations in the cell required for the onset of cancer is z – is as follows:

$$Dz\,(t) = -N_P\, e^{Nz\,(t)}\; Nz'(t) \tag{3.22}$$

Here the function $Nz'(t)$ is the time derivative of the function $Nz\,(t)$, and the function $Nz\,(t)$ is given by equations (3.17).

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 value of age distribution function $Dz\,(t)$ and the length of the interval dt.

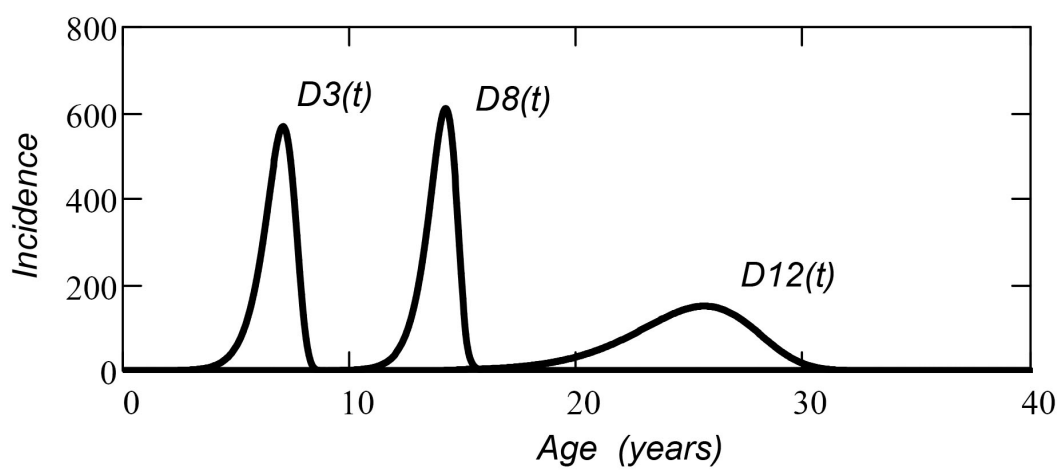

Figure 3.5. Age distribution of cancers in a group of N_P = 1000 people: with three mutations in the cell (the function $D3(t)$); with eight mutations in a cell (the function $D8(t)$); with twelve mutations in the cell (the function $D12(t)$).

The age distributions of cancers for cases where cancer occurs with three, eight and twelve mutations in the cell are shown in Fig. 3.5. The size of the group under consideration is $N_P = 1000$ people. We used the following parameters for plotting the graphs: $R_0 = 2.5 \cdot 10^7$; $k_0 = 0.5$; $k_1 = 1$; $k_3 = 0.44$; $t_1 = 15$; $a = 0.01$.

All obtained age distributions must satisfy the normalization condition – the area under the curve $Dz(t)$ is equal to the size of the group:

$$\int_0^\infty Dz(t)\,dt = N_P \qquad (3.23)$$

The age distribution can have a characteristic feature – a cusp (angle) in the function Dz (t) at the time t_1. The age distribution for nine and ten mutations in a cell is shown in Fig. 3.6. At the moment of time $t = 15$, the growth function of the cell population $N(t)$ enters the stage of limited growth, and the growth equation changes its form (see Equations 3.1).

The derivative of the growth function at this point is not a smooth function. Therefore, the age distribution can also have the noticeable cusp (angle).

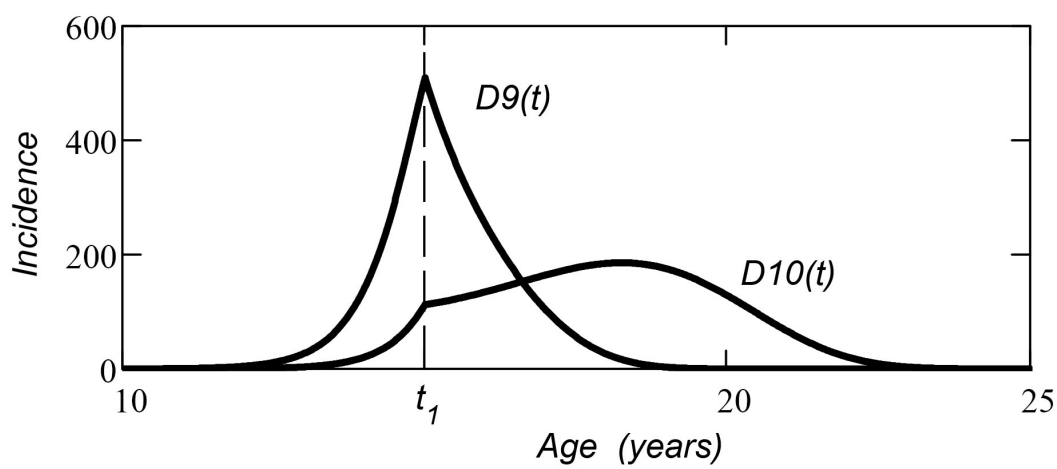

Figure 3.6. Age distribution of cancers in a group of $N_P = 1000$ people: with nine mutations in the cell (the function $D9(t)$); with ten mutations in a cell (the function $D10(t)$). The functions have a characteristic cusp at the time $t = 15$.

When we compare the age distributions presented in Fig. 3.5 with the age distributions obtained earlier for a simple mutation model (see Fig. 2.9), we see in fig. 3.5 curves are in the shape of a hill with a shallow (gentle) left slope and a steep right one. An exception is the case when the function reaches the top of the hill at time t_1 (see Fig. 3.6 and the function $D9(t)$).

In a simple mutational model (Fig. 2.9), the opposite is true – the right side of the hill is shallow.

We can get formulas (3.20) and (3.21) in another way. Perhaps some readers will understand this method more clearly.

Consider a group of N_P people. At the initial moment of time $t = 0$, the number of sick people is zero: $N3_P(0) = 0$. As before (for example), we believe that cancer in a cell begins when the cell receives three (or more) mutations.

The first cancer patient in the group will appear at the time t_1, when the function $N3_P(t)$ equals one.

$$N3_P(t_1) = N_P N3(t_1) = 1 \tag{3.24}$$

The $N3(t)$ function describes the number of cells in an anatomical organ that have three (or more) mutations.

After the first cancer patient appears, the number of healthy people in the group will decrease by one. We exclude sick people from consideration (they are already taken into account by statistics), despite the fact that they continue to receive mutations in their cells.

The second cancer patient in a group of $(N_P - 1)$ healthy people will appear at time t_2, when the condition is met:

$$\Big(N3(t_2) - N3(t_1)\Big)(N_P - 1) = 1$$

The number of sick people at the time t_2 is equal to two:

$$N3_P(t_2) = 2$$

The third cancer patient in a group of $(N_P - 2)$ healthy people will appear at time t_3, when the condition is met:

$$\Big(N3(t_3) - N3(t_2)\Big)(N_P - 2) = 1$$

The number of sick people at the time t_3 is equal to three:

$$N3_P(t_3) = 3$$

Proceeding in a similar way, we can obtain a discrete set of values of the function $N3_P(t_n) = n$ for a discrete set of time points t_n.

The general equation for finding t_n is as follows:

$$\Big(N3(t_n) - N3(t_{n-1})\Big)\Big(N_P - N3_P(t_{n-1})\Big) = 1 \tag{3.25}$$

Thus, we get the discrete function $N3_P(t_n)$ – this is a function that is only defined at some points. For practical use, it is more convenient to have a

continuous function, which can be obtained by considering the differential form of the equation:

$$dN3\,(t)\left(N_P - N3_P\,(t)\right) = dN3_P\,(t) \tag{3.26}$$

Here $dN3\,(t)$ and $dN3_P\,(t)$ are the increments of the functions $N3\,(t)$ and $N3_P\,(t)$ on a small time interval dt.

To solve the equation, we make the separation of variables:

$$dN3\,(t) = \frac{dN3_P\,(t)}{N_P - N3_P\,(t)} \tag{3.27}$$

and integrate both sides of the equation:

$$N3\,(t) + c_1 = -\,ln\left(N_P - N3_P\,(t)\right)$$

To find the constant c_1, we substitute the values of the functions at the time $t = 0$ into the equation:

$$c_1 = -\,ln\,(N_P)$$

So:

$$\frac{e^{N3\,(t)}}{N_P} = \frac{1}{N_P - N3_P\,(t)}$$

$$N3_P\,(t) = N_P\left(1 - e^{-N3\,(t)}\right)$$

We obtained formula (3.20) in a different way – we wrote and solved the differential equation (3.24). The formula describes the number of cancer patients $N3_P\,(t)$ in a group of N_P people, for the case when cancer begins with three mutations in the cell.

To obtain the age distribution $D3(t)$ of cancers in a group, it is necessary (as before) to find the derivative of the function $N3_P\,(t)$ (see Eq. 3.21).

The formula for the age distribution $Dz\,(t)$ can be written for any arbitrary function $N(t)$, which describes the growth of the cell population of anatomical tissue:

$$Dz(t) = \frac{d\left(N_P\,(1 - e^{-Nz\,(t)})\right)}{dt} = \frac{d\left(N_P\,(1 - e^{-N(t)\left(1-e^{-a\,t}\right)^z})\right)}{dt} \tag{3.28}$$

Here z is the number of cell mutations required for cancer to occur. The function $Nz\,(t)$ is:

$$Nz\,(t) = N\,(t)\left(1 - e^{-a\,t}\right)^z \tag{3.29}$$

4. Model of delayed carcinogenic event

The model of delayed carcinogenic event 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 (this 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 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 $N1(t)$:

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

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 $N1(t)$:

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

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

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

In this distribution, the age of T_S (age of onset carcinogenic effect) is the same for all people. This is not the case in real life. Some people have a slightly longer life expectancy than 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 $G1\,(t)$ over the entire group of people (with the standard deviation parameter σ):

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

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 $T_{S\,i}$.

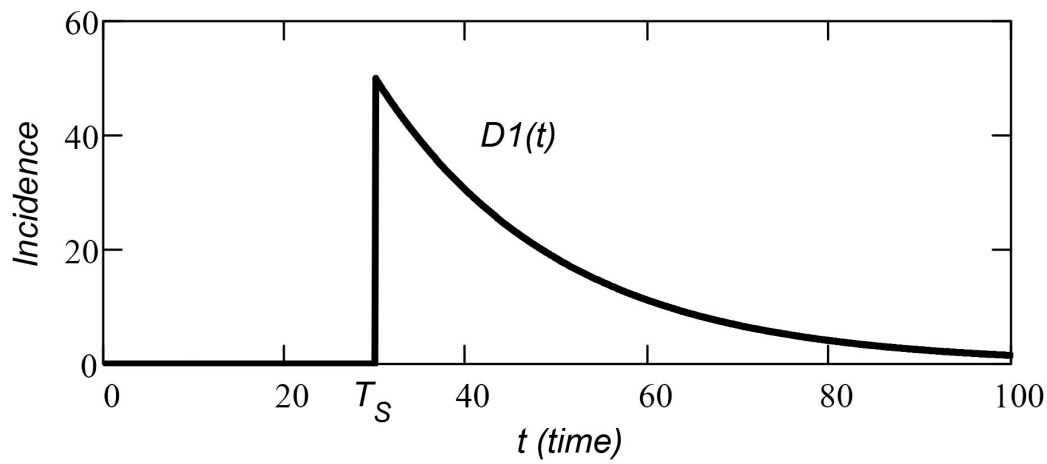

Figure 4.1. Age distribution of cancers for a group of 1000 people (with the time shift $T_S = 30$).

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 $t = dt$:

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

At the moment $t = 2dt$:

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

etc.

When dt tends to zero, we get:

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

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 function $D2\,(t)$ (age distribution of cancers in a group of 1000 people) is shown in Fig. 4.2.

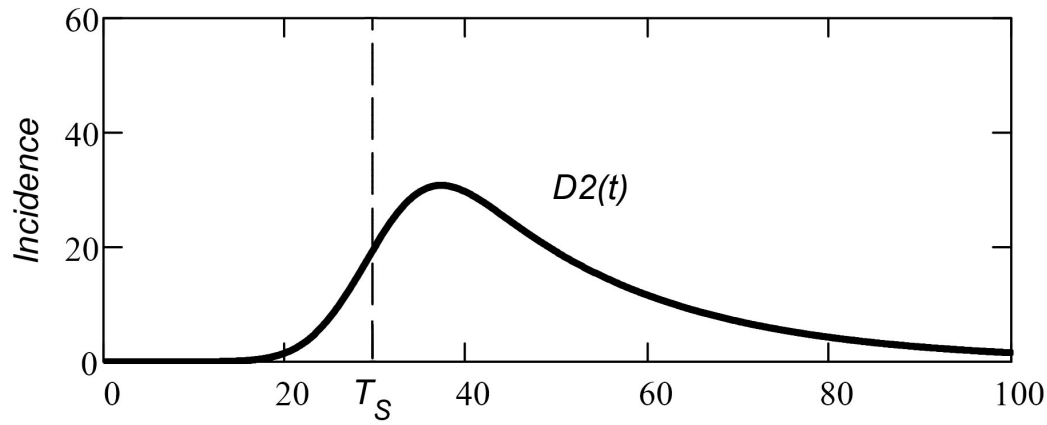

Figure 4.2. Function $D2(t)$ (age distribution of cancers in a group of 1000 people). Age shift is $Ts = 30$.

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 distributions of the incidence of lung, stomach, colon and breast (mammary) cancer in women and the age distributions of lung, stomach, colon and prostate cancer in men.

Age distribution data are given per 100,000 people – this 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, the graphs in the right column show data for male cancers.

6. Statistical analysis based on a simple mutational model

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

As an approximating function, we use the age distribution function $Dz(t)$ (See. Eq. 2.36):

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

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_P, and 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:

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

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):

$$Er_{\%}(z, a, N_0, T_S) = \frac{100}{n}\sqrt{\sum_{k=1}^{n}\left(\frac{D_k - Dz(t_k)}{D_{max}}\right)^2} \tag{6.3}$$

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_P 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.

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.

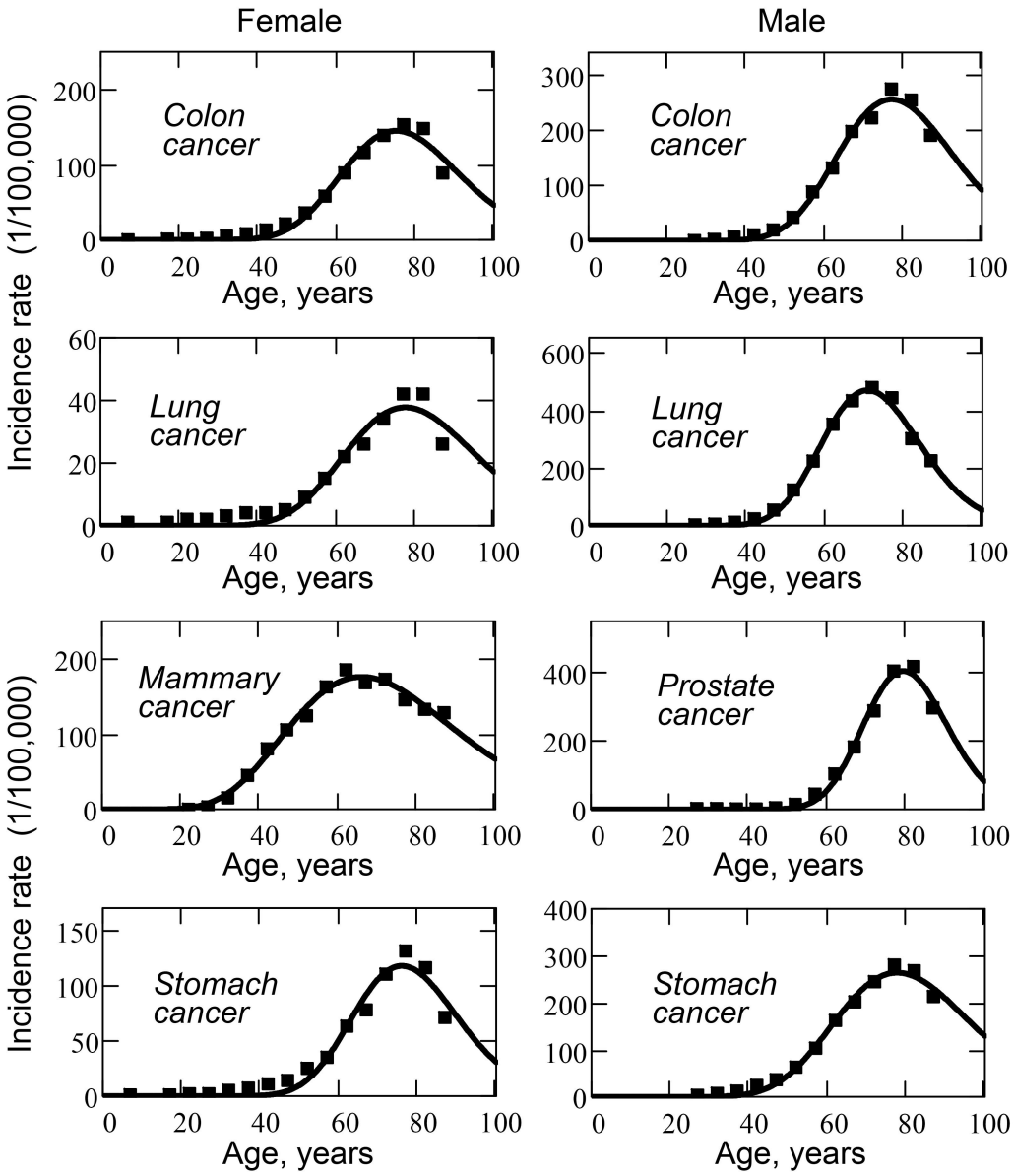

Figure 6.1. Age distributions of cancers per 100,000 people in women (left column) and men (right column), which are approximated by the function $Dz(t)$ within a simple mutational model.

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.

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.

Table 6.1.
Fitting parameters for a simple mutation model.

Female	z	T_S	a	N_P	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_P	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**

7. Statistical analysis based on a complex mutational model

A complex mutational model takes into account the growth of the cell population and the transfer of mutations obtained by the parent cells to the daughter cells. Cell population growth occurs in two stages: at the initial stage the population 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 $t = infty$.

We can estimate $N_G = R_0/k_0$ (the number of cells in a 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. In our case, we believe that growth rate is maximum at the age of 15 years ($t_1 = 15$). We use the following coefficient values: $k_0 = 0.5$, $k_1 = 1$. 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. The k_3 parameter is calculated from the condition $t_1 = 15$.

For a complex mutational model, we use the function $Dz(t)$ as an approximation function (fit function) for the age distribution of cancers (see Eq. 3.22):

$$Dz(t) = -N_P\, e^{Nz(t-T_S)}\; Nz'(t - T_S)$$

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 a time T_S since the onset of the disease).

That is, the T_S time must not be zero or negative.

We optimize the function by four parameters:z, a, N_P and 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:

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

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):

$$Er_{\%}(z, a, N_P, T_S) = \frac{100}{n}\sqrt{\sum_{k=1}^{n}\left(\frac{D_k - Dz(t_k)}{D_{max}}\right)^2} \qquad (7.1)$$

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).

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

Table 7.1.
Fitting parameters for a complex mutational model.

Female	z	T_S	a	N_P	Er	$Er_{\%}$
Colon	7	0.66	0.00087	4,729	1.02	0.67
Lung	6	8.54	0.00027	1,290	0.57	1.35
Mammary	4	3.71	0.00017	8,362	2.82	1.52
Stomach	8	1.10	0.00164	3,485	1.89	2.48
Averaged	**6.3**	**3.50**	**0.00074**	**4,466**	**1.57**	**1.51**

Male	z	T_S	a	N_P	Er	$Er_{\%}$
Colon	6	13.75	0.00090	8,032	1.83	0.67
Lung	6	10.39	0.00030	13,922	4.89	1.02
Prostate	12	0.76	0.00501	9,533	0.98	0.23
Stomach	6	6.67	0.00082	9,382	1.16	0.42
Averaged	**7.5**	**1.67**	**0.00176**	**10,217.2**	**2.22**	**0.59**

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

The average number of mutations required to form a cancer is 6.3 mutations per cell for female cancers and 7.5 mutations per cell for male cancers. The average number of mutations per unit time is $7 \cdot 10^{-4}$ mutations per year per cell for women and $1.8 \cdot 10^{-3}$ 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.

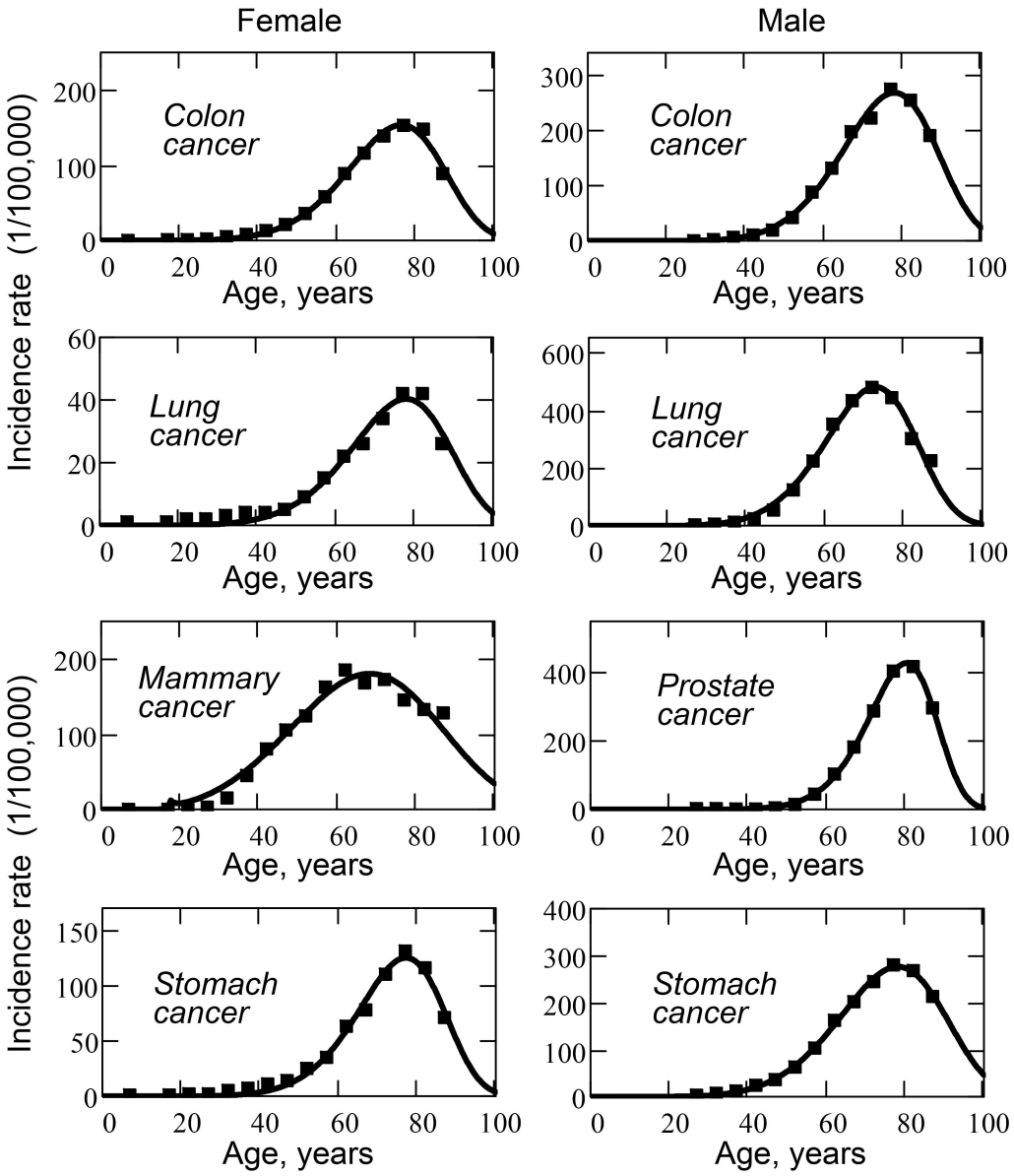

Figure 7.1. Age distributions of cancers in a group of 100,000 people in women (left column) and men (right column), which are approximated by the function $Dz(t)$ within a complex mutational model.

8. Statistical analysis based on model of delayed carcinogenic event

In the model of delayed carcinogenic event (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 is the same variable. 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:

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

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):

$$Er_{\%}(a, N_P, T_S, \sigma) = \frac{100}{n} \sqrt{\sum_{k=1}^{n} \left(\frac{D_k - D2(t_k)}{D_{max}} \right)^2} \qquad (8.2)$$

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, carcinogenic effect is directed not at an individual cell, but at the whole anatomical organ (at individual person). The average (for all types of cancer) number of people in the group forming a given age distribution of cancers is 6,860 people for women and 11,558 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.

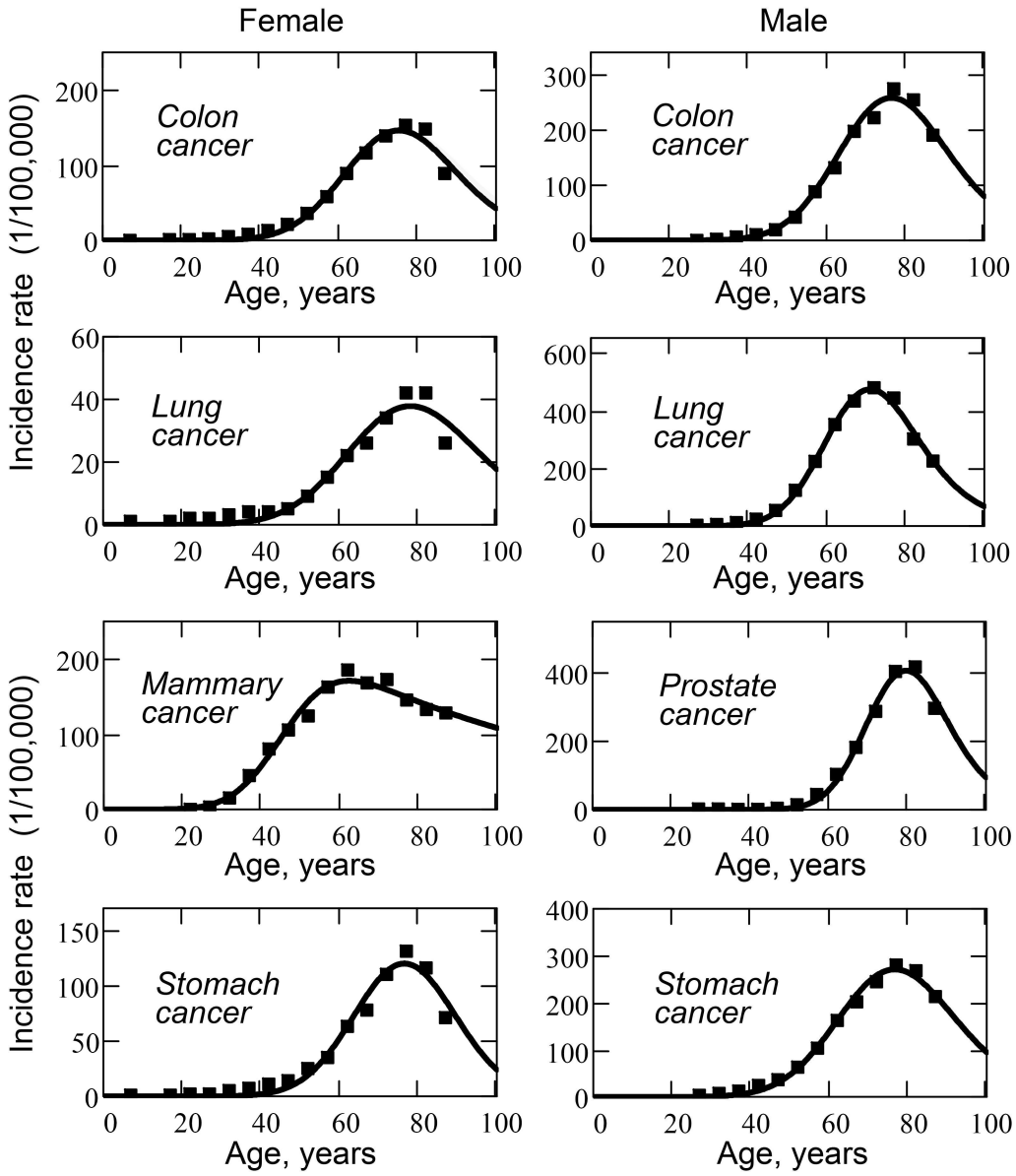

Figure 8.1. Age distributions of cancers in a group of 100,000 people in women (left column) and men (right column), which are approximated by the function $D2(t)$ within a model of delayed carcinogenic event (see Eq. 4.4).

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.

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

Female	N_P	T_S	a	σ	Er	$Er_{\%}$
Colon	5,430	67.8	0.10	12.3	1.82	1.19
Lung	1,610	69.0	0.08	13.8	0.73	1.74
Mammary	16,500	46.4	0.01	10.6	2.13	1.15
Stomach	3,900	72.9	0.25	12.5	1.77	1.35
Averaged	**6,860**	**64.0**	**0.11**	**12.3**	**1.61**	**1.36**

Male	N_P	T_S	a	σ	Er	$Er_{\%}$
Colon	9,480	69.2	0.09	11.6	2.48	0.91
Lung	15,100	63.9	0.10	10.0	3.11	0.65
Prostate	11,120	73.9	0.13	9.1	3.56	0.85
Stomach	10,530	69.0	0.10	13.2	2.32	0.83
Averaged	**11,558**	**69.0**	**0.11**	**11.0**	**2.87**	**0.81**

Part III
Discussion

9. Mutational models

Applying the mutational models considered in the work, we can determine (by applying the models to real age distributions) the following parameters of cancer:

– the number of mutations per cell required for cancer to appear (z parameter from Table 7.1);

– the average number of mutations per unit time (year) per one cell of the considered cell tissue (parameter a from Table 7.1);

– the time lag in medical diagnosis of cancer (the time interval between the occurrence of cancer in the cell and its detection), (T_S parameter from Table 7.1);

- the percentage of people susceptible (genetically predisposed) to this cancer (parameter $N_P/100,000$ from Table 7.1).

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 do know that the Erlang distribution describes a simplified mutational model that is not accurate enough to match reality.

In our case, in order to understand which of the models gives the best results for approximating the real age distributions of cancers, we can compare the parameters of the relative mean square error ($Er_{\%}$ parameters in Tables 6.1 and 7.1) for the mutational models presented in this work.

For female cancers, a simple mutation model gives the best results ($Er_{\%} = 1.46$). Although we know that a simple mutational model does not describe reality well. For male cancer, the complex model provides the best results ($Er_{\%} = 0.59$).

Another important parameter of the mutation model is the time lag between cancer formation and diagnosis.

However, when regression using a simple mutation model, this time delay is not detected (see T_S parameter in Table 6.1). Although this parameter is nonzero in the table, during optimization the program constantly reduces the time T_S to zero (and even to the area of negative values) in order to find the best approximation to a set of real data. Therefore, we write the last non-zero value of T_S into the table. Although, in theory, there should be a stable minimum of the error function to the left of the maximum of the practical data set. Medical practitioners usually estimate the diagnosis lag time (time from onset of cancer to diagnosis) at about 5-10 years.

In a complex mutational model, a delay in diagnosis is detected in only two (out of eight) cases in men – this is colon cancer and lung cancer, and the values $T_S = 13.75$ and $T_S = 10.39$, respectively (see table 7.1). The other six forms of cancer considered in the table (two male and four female cancers) do not have a time lag. As in the simple mutation model, we include in the table the last non-zero value of the T_S parameter, when the optimum of the error function was not found in the region of positive values of T_S.

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. This, perhaps, tells us that either the mutational hypothesis is wrong, or the presented mathematical model does not match reality well enough.

Quantitative analysis of mutational models shows results that are consistent with the well-established concepts of the mutational hypothesis of cancer formation. For example, for the female form of colon cancer, the size of the group of people providing the real age distribution of cancer is 4,729 people (for a complex mutational model). Although the real distribution is obtained for a group of 100,000 people. This may mean that only part of the group in question is predisposed to cancer (4,729 people of 100,000 – or 4.73%), and other people (94,271) do not develop cancer (probably for reasons of genetic resistance to the disease).

In other words, according to the data of the considered mutational model, cancer occurs only in a fraction of people who are probably genetically predisposed to the disease. (this is on average 4.47% of women and 10.22% of men for all types of cancer considered – see N_P parameter in Table 7.1). Cancer develops when one of the cells in the anatomical tissue receives a certain number of mutations. (on average, this is 6-7 mutations per cell – see z parameter in Table 7.1).

10. Model of delayed carcinogenic event

The delayed carcinogenic event model was developed as an alternative to the mutational models. Using this model, we can determine (by applying the model to real age distributions) the following parameters of cancer:

– the average number of carcinogenic events per unit time (year) per one considered anatomical organ (a parameter from Table 8.1);

– the percentage of people susceptible (genetically predisposed) to this cancer ($N_P/100,000$ parameter from Table 8.1).

– the sum of the time of the onset of carcinogenic effects and the time of delay in cancer diagnosis (T_S parameter from Table 8.1);

– the degree of non-simultaneity of the onset of carcinogenic effects in different people in the group (*sigma* parameter from Table 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).

The delayed carcinogenic event model (with the same number of parameters of the approximating function) gives a lower value of the relative mean square error ($Er_{\%}$ parameter) in comparison with both mutational models on average for all female forms of cancer (compare $Er_{\%}$ parameters in Tables 6.1, 7.1, and 8.1).

For male cancers, a complex mutational model is the best.

As well as for mutational models, we can quantify the main parameters of the mathematical model by comparing them with the real age distributions of cancers.

For example, for prostate cancer, the size of the group of people that provides the real age distribution of cancers is 11,120 people (men). This means that only a given number of men will get prostate cancer in their lives. This can be explained by a hereditary predisposition to the disease. That is, 11,56 % of men are prone to prostate cancer.

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). This is the a parameter from Table 8.1.

In any case, in order to further explore the models presented in this article, we need to match the models with a large body of real data on the age distribution of cancers.

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