A model for hepatocarcinogenesis with clonal expansion of three successive phenotypes of preneoplastic cells

Iris Geisler *, Annette Kopp-Schneider

Department of Biostatistics-R0700, German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Received 1 December 1999; received in revised form 22 August 2000; accepted 25 August 2000

Abstract

The two-stage model with clonal expansion of intermediate cells has often been used to describe the carcinogenesis process. The model hypothesizes that cells have to undergo two mutations on their way from the normal to the malignant stage. Biological experiments indicate the existence of three types of preneoplastic cells in hepatocarcinogenesis representing three successive intermediate stages in the development of malignant cells from normal cells. This finding suggests that hepatocarcinogenesis should be described by a multi-stage model with three intermediate stages, leading to a four-stage mutation model with clonal expansion of all types of intermediate cells. This model is presented and mathematical approximations for the number and size of nonextinct premalignant clones of the different cell types are derived. The model is applied to focal lesion data from a rat hepatocarcinogenesis experiment. © 2000 Elsevier Science Inc. All rights reserved.

Keywords: Multi-stage models; Clonal expansion; Hepatocarcinogenesis; Stereology

1. Introduction

There is a general consensus that the development of malignant tumors is a multi-stage process. It is supposed that cells have to undergo several changes on their way to malignancy. Armitage and Doll [1] hypothesized that a certain sequence of irreversible cell alterations has to be followed and that each cell moves from one stage to another independently of other cells. This model has been used frequently, although there was no cell proliferation incorporated.
Moolgavkar et al. [2] took up an idea of Kendall [3] and formulated a model including stochastic clonal expansion of normal and intermediate cells. For reasons of mathematical and analytical tractability, this model is restricted to one intermediate stage. They have shown that this model is consistent with data on one type of enzyme-altered foci (ATPase deficient foci) induced in a rat hepatocarcinogenesis experiment [4].

In another approach to analyze rat hepatocarcinogenesis experiments, phenotypic heterogeneity of the lesions has been observed (e.g., [5,6]). Three phenotypes of altered cell foci could be identified: clear cell foci, mixed cell foci and basophilic cell foci, representing successive stages in the development of liver carcinoma. Therefore, the application of a two-stage model for describing the hepatocarcinogenesis process seems inappropriate. Hence, an expanded multi-stage model taking into account the observed biological features of three intermediate cell types is called for.

In Section 2, the mathematical development for a model with clonal expansion of three types of intermediate cells is presented. Analytical expressions for the expected number and the size of nonextinct premalignant foci in an approximating model are derived. The application of the presented model and the classical two-stage model to hepatocarcinogenesis data from a study by Weber and Bannasch [5] is described in Section 3. In Section 4, the results and further research approaches are discussed.

2. Model description

The model describes the generation of three types of premalignant and of malignant cells. It is hypothesized that a normal cell is transformed stepwise into a malignant cell by irreversible changes to its genotype resulting in phenotypical alterations. The process starts with all normal cells. Once a normal cell undergoes a first critical mutation, an intermediate cell of type 1 is formed. Intermediate cells proliferate in a stochastic manner according to a birth–death process. If a normal cell sustains four mutations in sequence, a malignant cell will be generated.

**CELL DIVISION**

**DIFFERENTIATION - DEATH**

Fig. 1. A model with clonal expansion of three types of intermediate cells (solid lines) and possible extensions of the model incorporating clonal expansion of normal cells and malignant transformation (dashed lines).
The model is illustrated in detail in Fig. 1. Briefly, during a small time interval $\Delta t$, an intermediate cell of type $k$ ($k = 0, \ldots, 3$) may divide into two intermediate cells of type $k$ with probability $\beta_k(t) \Delta t + o(\Delta t)$, it may divide into one intermediate cell of type $k$ and one intermediate cell of type $k+1$ with probability $\mu_k(t) \beta_k(t) \Delta t + o(\Delta t)$, or it may differentiate or die with probability $\delta_k(t) \Delta t + o(\Delta t)$. The probability of more than one event occurring in time interval $[t, t + \Delta t]$ is equal to $o(\Delta t)$, and $o(\Delta t)$ is defined by $\lim_{\Delta t \to 0} (o(\Delta t)/\Delta t) = 0$.

The parameters $\beta_k$, $\mu_k$ and $\delta_k$ represent the birth, mutation, and death rates of type $k$ cells. It is assumed that all of these parameters are constant over time. The number of normal cells in a tissue is hypothesized to be a constant value over time, $x_0$, and therefore $\beta_0 = \delta_0$. Furthermore, it is supposed (at this point) that the proliferation behavior is the same in all intermediate stages. As in the two-stage model, an important characteristic of this model is the fact that every cell moves independently of other cells from one stage to the next.

Thus, it is assumed that
\[
\begin{align*}
\beta_1(t) &= \beta_2(t) = \beta_3(t) = \beta, \\
\delta_1(t) &= \delta_2(t) = \delta_3(t) = \delta, \\
\mu_0(t) &= \mu_1(t) = \mu_2(t) = \mu.
\end{align*}
\]

Let $I_k(t)$ be the number of intermediate cells of type $k$ at time $t$ and $I_0(t) \equiv x_0$ the constant number of normal cells. Using standard techniques for Markov processes, a partial differential equation for the probability generating function of the process can be derived (e.g. [7]), which can be used to obtain a system of ordinary differential equations for the expected numbers of intermediate cells
\[
\frac{\partial E[I_k(t)]}{\partial t} = \mu \beta E[I_{k-1}(t)] + (\beta - \delta)E[I_k(t)], \quad k \geq 1
\]
with initial conditions $E[I_0(0)] = x_0$ and $E[I_1(0)] = E[I_2(0)] = E[I_3(0)] = 0$. Eq. (1) is satisfied by
\[
E[I_k(t)] = \frac{x_0 \mu^k \beta^{k-1}}{(\beta - \delta)^k} e^{\mu \beta t} \left[ \sum_{i=1}^{k-1} (-1)^{3k-1-i} \frac{t(\beta - \delta)^i}{i!} + (-1)^{k+1}(1 - e^{-(\beta - \delta) t}) \right]
\]
for $\beta \neq \delta$ and by
\[
E[I_k(t)] = \frac{1}{k!} x_0 \beta_0 (\mu t)^k \beta^{k-1} \quad \text{for} \quad \beta = \delta.
\]

2.1. Approximation of the process

The formulation of the model with three types of intermediate cells implies that a later stage focus always arises within an earlier stage focus, e.g., a focus of type 3 is found in a focus of type 2 which again is contained in a focus of type 1, at least if the earlier stage foci are not extinct. Modeling this structure, i.e., modeling the joint likelihood of foci within foci is complicated. An attempt to derive the probability distribution for two types of cells (intermediate and malignant cells in the two-stage model) has been made by Dewanjii et al. [8]. However, the application of their results to the analysis of real data presents numerical problems.
The structure of the data necessitates the derivation of the marginal distributions for the size and the number of foci of different types. An approximation to these quantities is therefore needed. In the process as it is described in the previous section, type $k$ cells are generated from type $k-1$ cells with rate $\mu_{k-1} \beta_{k-1}$, $I_{k-1}(t)$ cells being at risk for transformation into type $k$ cells.

Using the same arguments as Sherman et al. [9] to approximate the number of malignant transformations from intermediate cells in the two-stage model, the process of generation of type $k$ cells from type $k-1$ cells can be approximated by a Poisson process with rate $\mu_{k-1} \beta_{k-1} E[I_{k-1}(t)]$. Hence in the approximate process, the rate of transformation into type $k$ cells does not depend on the actual number of type $k-1$ cells as it is the case in the original process, but it depends on their expected number. This approach has first been suggested by Whittemore and Keller [10] to approximate the distribution function for the time to first malignant transformation.

It should be noted that the expected number of cells of each type remains unchanged by the approximation. The approximate process can be understood as a multi-path two-stage process in which type $k$ cells are generated from normal cells with rate $\mu_{k-1} \beta_{k-1} E[I_{k-1}(t)]$, $E[I_{k-1}(t)]$ recursively depending upon the transformation rates and the birth and death rates of earlier stage cells.

### 2.2. Expected number of nonextinct premalignant foci

Let $N_k(t)$ be the number of nonextinct type $k$ foci at time $t$. Define the counting process for the number of non-extinct clones

$$V(t,s) = \begin{cases} 
0, & \text{if the birth-death process with birth rate } \beta \\
& \text{and death-rate } \delta \text{ starting at time } s \\
& \text{become extinct by time } t, \\
1, & \text{otherwise.}
\end{cases}$$

$N_k(t)$ can be written as filtered Poisson process (e.g. [11])

$$N_k(t) = \sum_{m=1}^{M_k(t)} V(t,s_m),$$

where $M_k(t)$ represents the number of mutational events in type $k-1$ cells leading to the formation of a type $k$ cell up to the time $t$ and $s_m$ denotes the times of these mutations and hence the time of formation of the resulting $N_k(t)$ foci. Given that $M_k(t) = n$, the generation times of the resulting type $k$ foci are independently and identically distributed with density function (e.g. [12])

$$\frac{E[I_{k-1}(s)]}{\int_0^t E[I_{k-1}(s)] \, ds} \text{ for } s \leq t.$$ 

Thus, the probability generating function (pgf) of the number of type $k$ foci is given by

$$\psi_{N_k}(u) = E[u^{N_k}] = \sum_{n=0}^{\infty} \left\{ E[u^{\sum_{m=1}^{n} V(t,s_m)}] \bigg| M_k(t) = n \right\} P(M_k(t) = n). \tag{4}$$

Due to the independence and the identical distribution of the generation times, the summands of Eq. (4) can be written as
Inserting Eq. (5) in Eq. (4) yields

\[
E \left[ u^{\sum_{m=1}^{n} V(t,s_m)} \mid M_k(t) = n \right] = \int_0^t \ldots \int_0^t \prod_{m=1}^{n} \left( \psi_{V(t,s_m)}(u) \frac{E[I_{k-1}(s_m)]}{\int_0^t E[I_{k-1}(s)] \, \mathrm{d}s} \right) \, \mathrm{d}s_1 \ldots \mathrm{d}s_n
\]

\[
= \left[ \frac{1}{\int_0^t E[I_{k-1}(s)] \, \mathrm{d}s} \int_0^t \psi_{V(t,s)}(u) E[I_{k-1}(s)] \, \mathrm{d}s \right]^n.
\]

(5)

Inserting Eq. (5) in Eq. (4) yields

\[
\psi_{N_k}(u) = \exp \left( \mu \beta \int_0^t E[I_{k-1}(s)] (\psi_{V(t,s)}(u) - 1) \, \mathrm{d}s \right).
\]

(6)

Let \( p_0(t,s) \) be the probability that a focus generated at time \( s \) becomes extinct by time \( t \), i.e., the extinction probability of a birth–death process starting in \( s \). This quantity is given by

\[
p_0(t,s) = \begin{cases} 
\frac{\delta(1 - e^{-(\beta-\delta)(t-s)})}{\beta - \delta e^{-(\beta-\delta)(t-s)}} & \text{for } \beta \neq \delta, \\
\frac{\beta(t-s)}{1 + \beta(t-s)} & \text{for } \beta = \delta.
\end{cases}
\]

(7)

Due to homogeneity of the birth–death process, \( p_0(t,s) = p_0(t-s,0) \). A comprehensive derivation of formula (7) can be found in a paper by Dewanji et al. [13]. Due to the construction of \( V(t,s) \),

\[
\psi_{V(t,s)}(u) = E[u^{V(t,s)}] = p_0(t,s) + u(1-p_0(t,s))
\]

and hence, Eq. (6) can be rewritten as

\[
\psi_{N_k}(u) = \exp \left( - (1-u) \mu \beta \int_0^t E[I_{k-1}(s)] [1 - p_0(t,s)] \, \mathrm{d}s \right),
\]

which corresponds to the pgf of a Poisson distributed random variable with parameter

\[
\mu \beta \int_0^t E[I_{k-1}(s)] [1 - p_0(t,s)] \, \mathrm{d}s.
\]

Thus, the expected number of nonextinct premalignant foci of type \( k \) is given by

\[
E[N_k] = \mu \beta \int_0^t E[I_{k-1}(s)] [1 - p_0(t,s)] \, \mathrm{d}s
\]

and hence for \( k = 1 \),

\[
E[N_1(t)] = \frac{x_0 \mu \beta_0}{\beta} [t(\beta - \delta) - \ln (1 - p_0(t,0))].
\]

(8)

For \( k > 1 \), a general formula can be obtained for \( \beta \neq \delta \)
\[ E[N_k(t)] = \mu \beta \int_0^t E[I_{k-1}(s)](1 - p_0(t, s)) \, ds \]
\[ = x_0 \beta_0 t^k \beta^{k-1} \left\{ \frac{(-1)^k}{\beta} \left[ \left( \frac{\delta - \beta e^{(\beta-\delta)t}}{\delta} \right) \ln(1 - p_0(t, 0)) - t(\beta - \delta) \right] \right. \]
\[ + \frac{e^{(\beta-\delta)t}}{\delta} \sum_{i=1}^{k-2} (1 - i)^{3k-i+1} (t(\beta - \delta))^i \frac{i!}{\max\{\beta, \delta\}} \left[ \ln \left( \frac{\max\{\beta, \delta\}}{|\beta - \delta|} \right) + I_{[\beta < \delta]} \frac{t(\beta - \delta)}{i + 1} \right] \]
\[ + \sum_{j=1}^{\infty} \min\{\beta, \delta\}^j \frac{e^{-(\beta-\delta)^j}}{j} \frac{1}{\max\{\beta, \delta\}} \right\} \right. \}
\[ = \frac{x_0 \beta_0 t^k \beta^{k-1}}{k!(1 + \beta t)} \Gamma_2\Gamma_1\left( k, 1, k + 1; \frac{\beta t}{1 + \beta t} \right) \] (see Appendix A)
\[ = \frac{x_0 \beta_0 t^k (1 + \beta t)^{k-1}}{(k - 1)! \beta} \ln(1 + \beta t) - \sum_{i=1}^{k-1} \frac{1}{s} \left( \frac{\beta t}{1 + \beta t} \right)^s. \]  

(9)

where
\[ I_{[\beta < \delta]} = \begin{cases} 1 & \text{for } \beta < \delta, \\ 0 & \text{for } \beta \geq \delta, \end{cases} \]  

(10)
denotes the indicator function and \( \Gamma_1(a, c; x) \) represents the confluent hypergeometric function (see Appendix A).

The case that \( \beta = \delta \) has to be treated separately and yields the formula
\[ E[N_k(t)] = \mu \beta \int_0^t E[I_{k-1}(s)](1 - p_0(t, s)) \, ds. \]
\[ = \frac{x_0 \beta_0 t^k \beta^{k-1}}{k!(1 + \beta t)} \Gamma_2\Gamma_1\left( k, 1, k + 1; \frac{\beta t}{1 + \beta t} \right) \] (see Appendix A)
\[ = \frac{x_0 \beta_0 t^k (1 + \beta t)^{k-1}}{(k - 1)! \beta} \ln(1 + \beta t) - \sum_{i=1}^{k-1} \frac{1}{s} \left( \frac{\beta t}{1 + \beta t} \right)^s. \]  

(11)

2.3. Size distribution of non-extinct premalignant foci

The size distribution \( S_k(t) \) of a nonextinct premalignant focus of type \( k \) can be expressed by the conditional size distribution given the generation time \( s \) of the focus
\[ P(S_k(t) = m \mid S_k(t) > 0) \]
\[ = \int_0^t P(S_k(t, s) = m \mid S_k(t, s) > 0) \frac{E[I_{k-1}(s)](1 - p_0(t, s))}{\int_0^t E[I_{k-1}(s)](1 - p_0(t, s)) \, ds} \, ds \]
\[ = \frac{\mu \beta}{E[N_k(t)]} \int_0^t P(S_k(t, s) = m \mid S_k(t, s) > 0) E[I_{k-1}(s)](1 - p_0(t, s)) \, ds, \]  

(12)

where \( m \) denotes the number of cells.

Note that the proliferation behavior is assumed to be the same in all intermediate stages and that the cells are subject to a homogeneous birth and death process. It follows that the conditional size distribution given the generation time depends only on the time period \( (t - s) \) and is independent of the type of the focus and of \( M_k(t) \), the number of mutational events in type \( k - 1 \) cells.
Consequently, the size distribution of foci in the two-stage model derived by Dewanji et al. [13] can be adopted. Conditional on a focus starting at time $s$ and not being extinct by time $t$, the distribution of its size is geometric with parameter $((\beta/\delta)p_0(t,s))$ ($\approx$ probability of success). Hence,

$$P(S_k(t,s) = m \mid S_k(t,s) > 0) = \left(\frac{\beta}{\delta}p_0(t,s)\right)^{m-1} \left(1 - \frac{\beta}{\delta}p_0(t,s)\right).$$

(13)

Inserting Eq. (13) in Eq. (12) yields

$$P(S_k(t) = m \mid S_k(t) > 0) = \frac{\mu \beta}{E[N_k(t)]} \int_0^t \left(\frac{\beta}{\delta}p_0(t,s)\right)^{m-1} \left(1 - \frac{\beta}{\delta}p_0(t,s)\right)(1 - p_0(t,s)) \, ds. \tag{14}$$

Explicit formulae for the size distribution of each foci type can be given by integration in (14):

- **Foci of type 1**

$$P(S_1(t) = m \mid S_1(t) > 0) = \frac{x_0 \mu \beta_0}{E[N_1(t)]} \int_0^t \left(\frac{\beta}{\delta}p_0(t,s)\right)^{m-1} \left(1 - \frac{\beta}{\delta}p_0(t,s)\right)(1 - p_0(t,s)) \, ds$$

$$= -\frac{1}{m} \left(\frac{\beta}{\delta}p_0(t,0)\right)^m \frac{1}{\ln (1 - p_0(t,0)) - (\beta - \delta)t}. \tag{15}$$

- **Foci of type 2**

$$P(S_2(t) = m \mid S_2(t) > 0) = \frac{\mu \beta}{E[N_2(t)]} \int_0^t \left(\frac{\beta}{\delta}p_0(t,s)\right)^{m-1} \left(1 - \frac{\beta}{\delta}p_0(t,s)\right)(1 - p_0(t,s)) \, ds$$

$$= \frac{x_0 \mu^2 \beta_0}{(\beta - \delta)E[N_2(t)]} \left(\frac{\beta}{\delta}\right)^{m-1} \left\{ -p_0(t,0)^m \left(1 - \frac{e^{(\beta-\delta)t}}{m \delta}\right) + \frac{(\beta - \delta)e^{(\beta-\delta)t}}{\delta^2} \right. \right.$$  

$$\left. \times \left[ \sum_{i=1}^{m} \frac{(p_0(t,0))^i}{i} + \ln (1 - p_0(t,0)) \right] \right\} \text{ for } \beta \neq \delta, \tag{16}$$

and

$$P(S_2(t) = m \mid S_2(t) > 0) = \frac{x_0 \mu^2 \beta_0}{\beta E[N_2(t)]} \left\{ \beta t(p_0(t,0))^m - \sum_{i=m+1}^{\infty} \frac{(p_0(t,0))^i}{i} \right\} \text{ for } \beta = \delta. \tag{17}$$

- **Foci of type 3**

$$P(S_3(t) = m \mid S_3(t) > 0) = \frac{\mu \beta}{E[N_3(t)]} \int_0^t \left(\frac{\beta}{\delta}p_0(t,s)\right)^{m-1} \left(1 - \frac{\beta}{\delta}p_0(t,s)\right)(1 - p_0(t,s)) \, ds$$

$$= \frac{x_0 \mu^3 \beta_0^2}{(\beta - \delta)^2E[N_3(t)]} \left(\frac{\beta - \delta)e^{(\beta-\delta)t}}{\delta^2}\left(\frac{\beta}{\delta}\right)^{m-1} \left[ -\frac{1}{i} \sum_{k=1}^{m} \left(\frac{j}{k}p_0(t,0)\right)^k \right. \right.$$  

$$\left. - \sum_{i=1}^{m} \frac{1}{i} \left(\frac{\delta}{\beta}\right)^i (\ln (1 - p_0(t,0)) - t(\beta - \delta)) + \sum_{i=1}^{m} \frac{1}{i} \sum_{k=1}^{i-1} \frac{(p_0(t,0))^k}{k} \right\} \text{ for } \beta \neq \delta, \tag{18}$$

and

$$P(S_3(t) = m \mid S_3(t) > 0) = \frac{x_0 \mu^3 \beta_0^2}{\beta E[N_3(t)]} \left\{ \beta t(p_0(t,0))^m - \sum_{i=m+1}^{\infty} \frac{(p_0(t,0))^i}{i} \right\} \text{ for } \beta = \delta. \tag{19}$$

\[
- \frac{\sum_{i=1}^{m} (p_0(t,0))^i}{i} + \ln(1 - p_0(t,0)) \left( \sum_{i=1}^{m} \frac{1}{i} - 1 \right) + (\beta - \delta) t \ln \left( \frac{|\beta - \delta|}{\max\{\beta, \delta\}} \right) \\
+ \text{Li}_2 \left( \frac{\min\{\beta, \delta\}}{\max\{\beta, \delta\}} \right) - \text{Li}_2 \left( \frac{\min\{\beta, \delta\} e^{-(\beta-\delta)t}}{\max\{\beta, \delta\}} \right) - I_{|\beta<\delta|} \frac{1}{2} (t(\beta - \delta))^2 \\
+ \frac{e^{(\beta-\delta)t}}{\beta m} \left[ \left( \frac{\beta}{\delta} \right)^m \sum_{i=1}^{m-1} \frac{(p_0(t,0))^i}{i} + \ln(1 - p_0(t,0)) - p_0(t,0)^m(1 - e^{-(\beta-\delta)t}) \right] \\
- \frac{m-1}{i} \left( \frac{\beta}{\delta} p_0(t,0)^i \right) - \ln(1 - p_0(t,0)) + t(\beta - \delta) \right] \} \quad \text{for} \quad \beta \neq \delta, \quad (18)
\]

where \(I_{|\beta|}\) denotes the indicator function defined in Eq. (10) and \(\text{Li}_2(x)\) is the dilogarithm function (see Appendix A).

- For \(\beta = \delta\):

\[
P(S_3(t) = m \mid S_3(t) > 0) = \frac{x_0 \mu^3 \beta_0}{2E[N_3(t)]} \left\{ \beta t^2 + 2t - \frac{2(1 + \beta t)}{\beta} \ln(1 + \beta t) + \frac{2}{\beta} \sum_{i=1}^{m-1} \sum_{k=1}^{i-1} \frac{(\beta t)^{k+1}}{k(k+1)(k+2)(1+\beta t)^k} \\
- (m-1) \left[ \frac{\beta t^2}{2} - t + \frac{\ln(1 + \beta t)}{\beta} \right] \right\}. \quad (19)
\]

### 2.4. Likelihood construction

The presented model describes the carcinogenesis process in terms of number and size of foci, the size of a focus being determined by its number of cells. The model can be used to describe the process in a three-dimensional situation, assuming that the foci are randomly allocated in the liver, and that foci are spherical with volume determined by the number of cells contained in the focus times the volume of a single cell. In practice, however, three-dimensional data are not available for hepatocarcinogenesis experiments. Instead, only the number and size of premalignant foci in a cross-sectional area of the liver can be observed. Thus, the number and size distributions of foci in the three-dimensional model derived in the previous sections have to be translated into the two-dimensional situation using a formula proposed by Wicksell [14], following the procedure proposed by Moolgavkar et al. [4]. Another feature which has to be included into the likelihood is the fact that due to the limits of the examination instruments, only foci greater than a critical size can be identified.

Let \(r_c\) be the radius of a single cell. The size of a focus expressed by the number of cells can be translated into the radius using the relationship \(m = (r/r_c)^3\), if cells are assumed to be tightly packed spheres. Therefore, the size distribution of the three-dimensional radii of type \(k\) foci can be written as
Further mathematical details about the Wicksell transformation can be found in the paper by Moolgavkar et al. [4]. Suppose that only foci with two-dimensional radii greater than \( \epsilon \) and lower than \( R \) can be observed.

The size distribution of the two-dimensional type \( k \) foci radii conditional on their being larger than \( \epsilon \) can be obtained by the Wicksell formula [14]:

\[
g_{2,k}(y) = \frac{y}{\mu'} \int_{y}^{\infty} \frac{1}{\sqrt{r^2 - y^2}} g_{3,k}(r) \, dr,
\]

where

\[
\mu' = \int_{\epsilon}^{\infty} \sqrt{r^2 - \epsilon^2} g_{3,k}(r) \, dr.
\]

The size distribution of the two-dimensional radii of type \( k \) foci conditional on their size being in \([\epsilon, R]\) is given by

\[
\tilde{g}_{2,k}(y) = \frac{g_{2,k}(y)}{G_{2,k}(R)}
\]

with the size distribution function of the radii of the foci in two dimensions

\[
G_{2,k}(y) = 1 - \frac{1}{\mu'} \int_{y}^{\infty} \sqrt{r^2 - y^2} g_{3,k}(r) \, dr.
\]

It can be shown that the observable number of foci in the cross-sectional area given their size lies in the interval \([\epsilon, R]\) is Poisson distributed with parameter

\[
2E[N_k(t)](G_{2,k}(R) - G_{2,k}(\epsilon))\bar{r}_{3,k},
\]

where

\[
G_{2,k}(y) = 1 - \frac{1}{\bar{r}_{3,k}} \int_{y}^{\infty} \sqrt{r^2 - y^2} g_{3,k}(r) \, dr
\]

represents the distribution function for the radii of the sections in two dimensions and

\[
\bar{r}_{3,k} = \int_{0}^{\infty} r g_{3,k}(r) \, dr
\]

denotes the mean radius in three dimensions.

Hence, the expected number (per area unit) of observable type \( k \) foci at time \( t \) in a cross-sectional area given their size lies in \([\epsilon, R]\) can be written as

\[
\tilde{n}_{2,k}(t) = 2E[N_k(t)] \left\{ \int_{\epsilon}^{\infty} \sqrt{r^2 - \epsilon^2} g_{3,k}(r) \, dr - \int_{R}^{\infty} \sqrt{r^2 - R^2} g_{3,k}(r) \, dr \right\}.
\]

Each animal evaluated at each time point contributes a data file containing information on
• the total area \( A \) of the examined transection,
• for each type, the number of observed foci with size in \([\epsilon, R]\), and
• the size in the two-dimensional transection of each of these foci.

Consider an animal evaluated at time point \( t \) with \( n_{2,k} \) observed foci of type \( k \) and \( r_{2,k,j} \) the observed two-dimensional radius of the \( j \)th focus of type \( k \). Then, the log-likelihood contribution of this animal is given by

\[
I(x_0\beta_0, \mu, \beta, \delta/\beta) = \sum_{k=1}^{3} \left[ \log \left( P( \text{the observed animal has } n_{2,k} \text{ foci of type } k) \right) + \sum_{j=1}^{n_{2,k}} \log \left( P( \text{the two-dimensional radius of focus } j \text{ of type } k \text{ is } r_{2,k,j}) \right) \right]
\]

\[
= \sum_{k=1}^{3} \left[ n_{2,k} \log(A\hat{n}_{2,k}) - A\hat{n}_{2,k} \right] + \sum_{j=1}^{n_{2,k}} \left[ \log(g_{2,k}^\epsilon(r_{2,k,j}) - \log(G_{2,k}^\epsilon(R)) \right]
\]

where \( C \) is a parameter independent constant depending upon the data. Each animal in the study adds an independent log-likelihood contribution. Due to the complexity of the likelihood, analytical expressions for the maximum likelihood estimates cannot be derived.

The log-likelihood depends on four model parameters: the product of the (constant) number of normal susceptible cells in the tissue of interest, \( x_0 \), and the birth rate of normal cells, \( \beta_0 \), (note that \( x_0 \) and \( \beta_0 \) are not separately identifiable), the mutation rate, \( \lambda \), the birth rate of intermediate cells, \( \beta \), and the ratio of the death and the birth rate, \( \delta/\beta \).

3. Application to liver focal lesion data

There is strong evidence for the carcinogenicity of N-nitrosomorpholine (NNM) in experimental animals (e.g., [5,15,16]). In these studies, male Sprague–Dawley rats were subjected to NNM in the drinking water in nine different treatment groups. The presented model is applied to the data of one of these exposure groups [5]. In this experiment, the rats were treated by continuous oral administration of 6 mg/kg body weight of NNM. At 11, 15, 20, 27 and 37 weeks, respectively, after the start of the exposition, five animals were killed. One liver slice per animal was stained by H&E. Preneoplastic foci of altered hepatocytes were classified into different phenotypes according to Bannasch and Zerban [17] and evaluated by morphometry. Alterations in enzyme expression of the foci were not considered in this study.

A normal hepatocyte is assumed to have a radius of about 0.012 mm. Lesions smaller than 0.003 mm\(^2\) could not be detected, so that the smallest observable radius \( \epsilon = 0.031 \) mm. No foci were excluded from the analysis and hence \( R = \infty \) in Eq. (20). The likelihood was maximized numerically using the optimizing software package MINUIT from the CERN Program Library. During the estimation procedure, numerical integration of various expressions was performed, using the Numerical Algorithms Groups (NAg) Fortran Library, Version Mark 18.
First, the three preneoplastic types of cells are united into one intermediate stage and a classical two-stage model is applied to the data. The four parameter estimates obtained by maximum likelihood estimation (expressed as daily rates) were the number of normal cells in cell division \( x_0 \beta_0 = 58431 \) (0.95%-CI [47312, 72163]) per day and cell, the mutation rate \( \mu = 4.971 \times 10^{-5} \) (0.95%-CI [4.025 \times 10^{-5}, 6.140 \times 10^{-5}]) per day and cell, the birth rate \( \beta = 3.910 \) (0.95%-CI [2.665, 5.737]) per day and cell, the ratio of the death and the birth rate \( \delta/\beta = 0.99455 \) (0.95%-CI [0.99222, 0.99688]), and the value of the log-likelihood was 10070.223 \( + C \) with a constant \( C \) depending upon the data. The parameter estimates coincide with the results published by Moolgavkar et al. [4]. The number of foci and the size distribution of the foci are well predicted by the model. However, experiments show that the different phenotypes of foci are distinct in a number of metabolic parameters important in the development of malignancy [18–20]. Therefore, only one intermediate stage collecting all foci regardless of type seems inappropriate to describe the hepatocarcinogenesis process.

The model presented in this manuscript incorporates the biological feature of three successive intermediate stages and is used to describe three types of foci: clear/acidophilic, mixed and basophilic cells. A mixed cell focus is characterized by the existence of both clear/acidophilic and basophilic cells in the focus, and hence, it can be considered to be a lesion in the process of transformation from a clear/acidophilic cell focus to a basophilic cell focus. Thus, a model with three successive types of intermediate cells may approximate the biological process. Biological studies [6] suggest that type 1 intermediate cells can be identified in clear/acidophilic foci, type 2 intermediate cells as the cells in a mixed focus and type 3 intermediate cells as those in a basophilic focus.

A variety of four-stage models was applied to the data. It should be stressed that all of these models are approximate models (cf. Sections 2.1 and 2.4). Maximum likelihood estimates can therefore be interpreted in the context of the approximate models only. The impact of the approximation on the bias of the parameter estimates cannot be assessed.

At first, the proliferation behavior was assumed to be the same in all stages. The four parameter estimates obtained by maximum likelihood estimation (expressed as daily rates) were the number of normal cells in cell division \( x_0 \beta_0 = 1047100 \) (0.95%-CI [876850, 1250300]) per day and cm\(^3\), the mutation rate \( \mu = 2.655 \times 10^{-5} \) (0.95%-CI [2.346 \times 10^{-5}, 3.004 \times 10^{-5}]) per day and cell, the birth rate \( \beta = 79.510 \) (0.95%-CI [79.509, 79.510]) per day and cell, the ratio of the death and the birth rate \( \delta/\beta = 0.99994 \) (0.95%-CI [0.99993, 0.99994]), and the value of the log-likelihood was 8640.413 \( + C \) with a constant \( C \) depending upon the data. Among the estimated parameters, the birth rate attracts attention. A birth rate \( \beta = 79.51 \) per day and cell corresponds to an average cell cycle time of about 18 min, which is biologically unrealistic.

Fig. 2 shows the expected in comparison with the observed number of all three types of foci in the two-dimensional situation with the parameters obtained from maximum likelihood estimation. The empirical and the theoretical size distribution function in the two-dimensional transection of the three types of foci are presented exemplary for \( t = 37 \) weeks in Fig. 3. It is obvious from the figures that the number of foci is well fitted by the model whereas the empirical and the theoretical size distribution function differ greatly. For type 1 foci, the theoretical size distribution function lies for the most part under the empirical size distribution function, indicating that radii of type 1 foci are overestimated by the model. For foci of type 2 and 3, the situation is inverted. In this case, the empirical size distribution function is lower than the theoretical one and therefore the radii of type 2 and 3 foci are underestimated by the model.
The assumption of constant proliferation rates in all stages may not be in accord with the biological process. Experiments show that cell proliferation increases as the foci are in a more advanced stage towards malignancy [21]. Therefore, in a second approach, the model was modified to allow for different birth rates for the intermediate stages in order to improve the fit of the model to the data.

The ratio of the death and the birth rate was kept constant for all types of intermediate cells. Maximum likelihood methods yielded the number of normal cells in cell division $n_0 b_0 = 45078 (0.95\%-CI [42038, 48338])$ per day and cm$^3$, the mutation rate $\hat{\mu} = 2.065 \times 10^{-4} (0.95\%-CI [1.966 \times 10^{-4}, 2.170 \times 10^{-4}])$ per day and cell, the birth rate of clear/acidophilic cells $\hat{b}_1 = 15.563 (0.95\%-CI [14.707, 16.469])$ per day and cell, the birth rate of mixed cells $\hat{b}_2 = 51.975 (0.95\%-CI [47.778, 56.540])$ per day and cell, the birth rate of basophilic cells $\hat{b}_3 = 1.05 \times 10^{-5} (0.95\%-CI [4.29 \times 10^{-6}, 2.6062 \times 10^{-5}])$ per day and cell and the ratio of the death and the birth rate $\hat{d}_k/\hat{b}_k = 0.99956 (0.95\%-CI [0.99954, 0.99959]) (k = 1, 2, 3)$. The value of the log-likelihood was $9009.806 + C$ showing that the fit of the model to the data could be improved compared to the analysis with equal cell division rates. Reported here are the parameter estimates of the global maximum of the log-likelihood. A local maximum very close to the global maximum can be found with parameter estimates very similar to those in the global maximum, the only difference being the estimate $\hat{b}_3 = 2.443 \times 10^5$.

The expected and the observed number of all three types of foci with the parameters obtained from maximum likelihood estimation are illustrated in Fig. 4. Fig. 5 shows the empirical and the theoretical size distribution function of clear/acidophilic and mixed cell foci exemplary for $t = 37$. Note that for basophilic cell foci no size distribution function can be given, because the expected number of basophilic cell foci at time $t = 37$ weeks is zero (see Fig. 4).

It can therefore be seen that relaxing the condition of constant growth rates in all stages does not amend the lack of fit of the model. Two of the estimated birth rates ($\hat{b}_1$ and $\hat{b}_2$) are much too
high and the third \( \beta_3 \) is much too low, and there is a tendency to underestimate radii of younger foci and overestimate radii of elder foci.

Investigating the reasons for the poor fit of the model, some of the assumptions regarding the size of foci in terms of number of cell were changed. Until now, foci were assumed to consist of tightly packed spherical cells without spatial gaps. Assuming that intermediate cells are contained in cubes, the number of cells in a focus of size \( r \) is given by the relationship \( m = (\pi/6)(r/r_c)^3 \). In addition, biological experiments [22] indicate that altered hepatocytes may be larger than normal

---

**Fig. 3.** Theoretical and empirical size distribution function of (a) clear/acidophilic, (b) mixed and (c) basophilic cell foci radii in the two-dimensional transection with the parameters obtained from maximum likelihood estimation of the four-stage model with equal birth rates in all stages as applied to the data of Weber and Bannasch [5]. Shown are the size distribution functions at time \( t = 37 \) weeks.
Fig. 4. Expected and observed number of (a) clear/acidophilic, (b) mixed and (c) basophilic cell foci in the two-dimensional transection with the parameters obtained from maximum likelihood estimation of the four-stage model with different birth rates in all stages as applied to the data of Weber and Bannasch [5].

Fig. 5. Theoretical and empirical size distribution function of (a) clear/acidophilic and (b) mixed cell foci radii in the two-dimensional transection with the parameters obtained from maximum likelihood estimation of the four-stage model with different birth rates in all stages as applied to the data of Weber and Bannasch [5]. Shown are the size distribution functions at time $t = 37$ weeks.
hepatocytes, with a radius of approximately $r_c = 0.014 \text{ mm}$. Both alterations of the model assumptions lead to a reduction in the number of intermediate cells contained in the foci compared to the former assumptions. With this model modification, the estimates for the birth rates are both in the model with identical birth rates for the different types of cells ($\hat{\beta} = 24.519$) and in the model with individual birth rates ($\hat{\beta}_1 = 7.893, \hat{\beta}_2 = 157.260, \hat{\beta}_3 = 2.625 \times 10^5$) still much too high. Again, the likelihood has a local maximum close to the global maximum with all parameter estimates similar, but $\hat{\beta}_3 = 4.290 \times 10^{-6}$, therefore again in an implausible range.

In another approach, the model was extended to incorporate not only NNM induced foci but also spontaneously developing foci. The time interval from birth to start of the NNM experiment was included in the present model. During this time period, foci develop with a background mutation rate $\gamma$, whereas foci are generated with mutation rate $\mu$ in addition to the background rate after the start of NNM treatment. Our data did not allow for the estimation of separate birth rates before and during treatment because the first sacrifice was at 11 weeks after the start of treatment. In the experiments described above, the rats were subjected for the first time to NNM at the age of 14 weeks. Incorporating background mutation before the start of the experiment and the changed assumptions for the number of intermediate cells in the foci described in the previous paragraph yields a local maximum close to a global maximum of the log-likelihood. In the local maximum, the parameter estimate for the birth rate ($\hat{\beta} = 0.292$) lies in the biological reference interval, but the estimate for the background rate ($\hat{\gamma} = 0.005$) is more than hundred times that of the estimate of the NNM induced mutation rate, which is an implausible parameter constellation. In the global maximum, the background and the NNM induced mutation rates are in realistic relationship, but the parameter estimate for the birth rate is again much too high ($\hat{\beta} = 24.098$). These facts substantiate the lack of fit of this modified model in particular regarding the size of the foci. For our data set, the mutational model can only fit the rare but large cell foci in later stages with an unrealistic high estimate for the birth rate or with an unrealistic high value for the background rate showing that in the model the foci are generated earlier than in reality.

4. Discussion and summary

A model with three successive types of preneoplastic cells is used to evaluate data on number and size of preneoplastic clones in a cross-sectional area of the rat liver. This model incorporates transformation of single cells to the next stage and clonal expansion of all three intermediate stages.

A basic assumption of the model is the sequence of phenotypical changes of the liver foci. Six different models can be formulated, one for each permutation of the three phenotypical classes. Each of the six models was applied to the data and the resulting six values of the maximum log-likelihood were compared (results not shown). Two sequences (clear/acidophilic → mixed → basophilic and basophilic → mixed → clear/acidophilic) are favored by the data. This result coincides with the morphological feature that a mixed cell focus consists of clear/acidophilic cells and basophilic cells, and hence shows a lesion in the process of transformation from a clear/acidophilic cell focus to a basophilic cell focus or from a basophilic cell focus to a clear/acidophilic cell focus.
Consider the sequence clear/acidophilic cell foci → mixed cell foci → basophilic cell foci. The results obtained by maximum likelihood methods give rise to some discussion. The parameter estimate for the birth rate is biologically not explainable. In addition, the radii estimated by the model do not approximate the observed situation. Radii of types 2 and 3 foci are underestimated by the model. The reason for this may be that the observed mean foci radius is increasing from clear/acidophilic cell foci to mixed cell foci and to basophilic cell foci [5]. For the model used here the expected foci radii are ordered just in the opposite direction since the further the focus is on its way to malignancy, the later it was generated and therefore the smaller it is. The model presented in this paper predicts that earlier stage clones are more frequent and larger than later stage clones. The observations made in the experiment indicate that earlier stage foci are smaller and more frequent than later stage clones (see Table 1).

Even if the order of the sequence in the model was reversed, resulting in the sequence basophilic → mixed → clear/acidophilic cell foci, it is not possible to reconcile the data with the model. In this case, the model predicts that basophilic cell foci are larger and more frequent than mixed cell foci and clear/acidophilic cell foci, whereas in the experiment basophilic cell foci are less frequent and larger than mixed cell foci and clear/acidophilic cell foci. Thus, the size of the foci would be well fitted by the model whereas the number of the earlier stage foci would be underestimated and the number of the later stage clones would be overestimated by the model (see Table 2).

It is obvious from the tables that the presented model is inappropriate to describe adequately both the number of foci and the size of foci in the hepatocarcinogenesis process. The question is how to model the development of rare but large cell foci from frequent and small cell foci.

Table 1
Qualitative comparison of model predictions and experimental observations of number and size of three premalignant foci for the sequence of phenotypical changes given by clear/acidophilic → mixed → basophilic

<table>
<thead>
<tr>
<th>Model</th>
<th>Clear/acidophilic</th>
<th>→</th>
<th>Mixed</th>
<th>→</th>
<th>Basophilic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of foci</td>
<td>Frequent</td>
<td>&gt;</td>
<td>⋯</td>
<td>&gt;</td>
<td>Rare</td>
</tr>
<tr>
<td>Size of foci</td>
<td>Large</td>
<td>&gt;</td>
<td>⋯</td>
<td>&gt;</td>
<td>Small</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Clear/acidophilic</th>
<th>→</th>
<th>Mixed</th>
<th>→</th>
<th>Basophilic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of foci</td>
<td>Frequent</td>
<td>&gt;</td>
<td>⋯</td>
<td>&gt;</td>
<td>Rare</td>
</tr>
<tr>
<td>Size of foci</td>
<td>Small</td>
<td>&lt;</td>
<td>⋯</td>
<td>&lt;</td>
<td>Large</td>
</tr>
</tbody>
</table>

Table 2
Qualitative comparison of model predictions and experimental observations of number and size of three premalignant foci for the sequence of phenotypical changes given by basophilic → mixed → clear/acidophilic

<table>
<thead>
<tr>
<th>Model</th>
<th>Basophilic</th>
<th>→</th>
<th>Mixed</th>
<th>→</th>
<th>Clear/acidophilic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of foci</td>
<td>Frequent</td>
<td>&gt;</td>
<td>⋯</td>
<td>&gt;</td>
<td>Rare</td>
</tr>
<tr>
<td>Size of foci</td>
<td>Large</td>
<td>&gt;</td>
<td>⋯</td>
<td>&gt;</td>
<td>Small</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Basophilic</th>
<th>→</th>
<th>Mixed</th>
<th>→</th>
<th>Clear/acidophilic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of foci</td>
<td>Rare</td>
<td>&lt;</td>
<td>⋯</td>
<td>&lt;</td>
<td>Frequent</td>
</tr>
<tr>
<td>Size of foci</td>
<td>Large</td>
<td>&gt;</td>
<td>⋯</td>
<td>&gt;</td>
<td>Small</td>
</tr>
</tbody>
</table>
A solution to this problem may be given by the so-called color-shift model presented by Kopp-Schneider et al. [23]. This model is founded on the assumption that preneoplastic foci irreversibly change their phenotype as an entity rather than by mutation of a single cell in the colony. The color-shift model was applied to the same data set described in the previous section, and it could be shown that the sequence of phenotypical changes of the foci suggested by biological studies was preferred by the model and it suggests that foci start from a large collection of cells when they are generated.

In the present manuscript it is shown that a model with clonal expansion of three successive intermediate cell types cannot be used to explain the data from this hepatocarcinogenesis experiment, therefore suggesting that the change of phenotype does not involve a mutational step but rather a type of field effect that transforms the whole focus to the next phenotype. This finding stresses the strength of formulating stochastic models and applying them with maximum likelihood methods to test biological hypotheses in a rational fashion.

Acknowledgements

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) No. KO 1886/1-1 to Iris Geisler. We thank Professor Peter Bannasch for letting us use the data, for frequent and fruitful discussions as well as for the help with the biological interpretation of our findings. We thank Dr Georg Luebeck for his comments on an earlier version of the manuscript.

Appendix A. Mathematical appendix

The hypergeometric function is defined by

$$
2F_1(a, b, c; x) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 y^{a-1}(1-y)^{c-a-1}(1-xy)^{-b} dy \quad (c > a > 0)
$$

$$
= \sum_{r=0}^{\infty} \frac{(a)_r(b)_r}{r!(c)_r} x^r
$$

with \((a)_0 = 1\) and \((a)_r = a(a+1)\cdots(a+r-1)\). This series is convergent for \(c > 0\) and \(|x| < 1\) (e.g. [24]).

For \(b = 1\) and \(c = a + 1\), the hypergeometric function can be written as

$$
2F_1(a, 1, a + 1; x) = \sum_{r=0}^{\infty} \frac{a}{a+r} x^r.
$$

The confluent hypergeometric function (Kummer-function) is defined by

$$
1F_1(a; c; x) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 y^{a-1}(1-y)^{c-a-1}e^{xy} dy \quad (c > a > 0)
$$

$$
= \sum_{r=0}^{\infty} \frac{(a)_r}{r!(c)_r} x^r.
$$

This series is convergent for all \(x\).
For \( c = a + 1 \), the formula reduces to

\[
\sum_{r=0}^{\infty} \frac{a}{(a+r)r!} x^r.
\]

The dilogarithm function is defined by

\[
\text{Li}_2(x) = - \int_0^x \frac{\ln(1-y)}{y} \, dy.
\]

This integral converges for \(| x | < 1\).

Of particular interest for our problem is the relationship between the dilogarithm function occurring in the size distribution function of type 3 foci and the confluent hypergeometric function which can be found in the formula for the expected number of non-extinct premalignant foci. Considering the difference of two dilogarithm functions for the situation

\[
\text{Li}_2\left(\frac{\delta}{\beta}\right) - \text{Li}_2\left(\frac{\delta}{\beta} e^{-(\beta-\delta)t}\right) = - \int_{\delta/\beta}^{\delta/\beta e^{-(\beta-\delta)t}} \frac{\ln(1-y)}{y} \, dy \quad \text{for} \quad \beta > \delta
\]

and substituting \( y = (\delta/\beta) e^{-(\beta-\delta)(t-s)} \) yields

\[
\text{Li}_2\left(\frac{\delta}{\beta}\right) - \text{Li}_2\left(\frac{\delta}{\beta} e^{-(\beta-\delta)t}\right) = - (\beta-\delta) \int_0^t \ln \left(1 - \frac{\delta}{\beta} e^{-(\beta-\delta)(t-s)}\right) \, ds
\]

\[
= t(\beta-\delta) \sum_{j=1}^{\infty} \left(\frac{\delta}{\beta}\right)^j \frac{e^{-(\beta-\delta)tj}}{j} \, _1\text{F}_1(1, 2; (\beta-\delta)jt).
\]

Similarly, it can be shown that for \( \beta < \delta \)

\[
\text{Li}_2\left(\frac{\beta}{\delta}\right) - \text{Li}_2\left(\frac{\beta}{\delta} e^{(\beta-\delta)t}\right) = t(\beta-\delta) \sum_{j=1}^{\infty} \left(\frac{\beta}{\delta}\right)^j \frac{e^{(\beta-\delta)tj}}{j} \, _1\text{F}_1(1, 2; (\beta-\delta)jt).
\]

References