Two stage model for carcinogenesis: number and size distributions of premalignant clones in longitudinal studies

A. Dewanji [a,b], M.J. Goddard [c], D. Krewski [c], S.H. Moolgavkar [a,*]

[a] *Fred Hutchinson Cancer Research Center, 1100 Fairiew Ave. N. MP-665, Seattle, WA 98109-1024, USA*

[b] *On leave from Indian Statistical Institute, Calcutta, India*

[c] *Health Canada, Ottawa, Ontario, Canada K1A 0L2*

Received 5 July 1995; received in revised form 3 August 1998; accepted 12 October 1998

Abstract

The two stage clonal expansion model of carcinogenesis provides a convenient biologically based framework for the description of toxicologic and epidemiologic data on carcinogenesis. Under this model, a cancer cell is generated following the occurrence of two critical mutations in a single stem cell. Initiated cells that have sustained the first mutation undergo a stochastic birth–death process resulting in clonal expansion of the initiated cell population. In this article, we consider the analysis of longitudinal data on the number and size of premalignant clones, formed by clonal expansion of initiated cells. In particular, the joint distribution of the number of premalignant clones observed at different points in time in the same subject is derived. The application of these results in the statistical analysis of longitudinal data on the number and size of premalignant clones observed in initiation-promotion experiments is indicated.

Keywords: Initiation; Promotion; Papillomas

1. Introduction

Armitage and Doll [1] pioneered the development of multistage models of carcinogenesis, in which a cancer cell is formed following the occurrence, in sequence, of a number of mutations considered to represent different stages of the process of malignant transformation. Subsequently, Kendall [9], Moolgavkar and Venzon [22] and Moolgavkar and Knudson [18] developed multistage models that explicitly considered cell proliferation kinetics. Moolgavkar–Venzon–Knudson noted that, providing cell proliferation was explicitly incorporated, a model postulating two rate limiting events on the pathway to malignancy was consistent with a wide range of toxicologic and epidemiologic data. This two stage model has been successfully used to describe both toxicologic and epidemiologic data on carcinogenesis [20,13]. For some variations of this model, see Refs. [23,11].

The mathematical aspects of the two stage model are described by Moolgavkar and Luebeck [20] and Tan [25]. Briefly, let $X(t)$ denote the number of susceptible stem cells in the tissue of interest at time t. The number of intermediate cells $I(t)$ that have sustained the first mutation by time t follows a Poisson process with intensity $v(t)X(t)$, where $v(t)$ denotes the first stage mutation rate. Initiated cells undergo a stochastic birth–death process, with $\alpha(t)$ and $\beta(t)$ denoting the birth and death (or differentiation) rates at time t. A malignant cancer cell is formed following the occurrence of a second mutation in one of the intermediate cells. Denes and Krewski [5] have recently developed an exact expression for the probability generating function of this two stage model allowing for stochastic stem cell growth. In practice, however, $X(t)$ is usually taken to be deterministic because the number of normal cells in a tissue is under tight homeostatic control.

The two stage model provides a rich biologically based framework for describing the process of carcinogenesis. In this paper, we use this model to describe the number and size of premalignant clones of intermediate cells [7]. Premalignant clones appear as enzyme altered liver foci in rodent hepatocarcinogenesis experiments [24] and as skin papillomas in initiation-promotion experiments on the mouse skin [4,2,17].

In the mouse skin model, it is possible to obtain information on the evolution of papillomas in the same subject over time. This article focuses on the use of the two stage model to describe longitudinal data on the number and size of premalignant lesions. Kopp-Schneider and Portier [10,12] have earlier used the two stage model and some modifications of it to analyze data on mouse skin papillomas; however, they ignored the inherent dependence present in the longitudinal observations made on the same subject over time. The purpose of this article is to express this dependence quantitatively within the framework of two stage model and indicate how to incorporate it in analyses. This article presents the mathematical results required for analyses of data on premalignant clones only. In Section 2, we derive the distributional results for the number of premalignant clones observed in the same subject at different points in time. Section 3 considers extension of the results for the case in which observations on the sizes of the premalignant clones, in addition to the numbers, are available. The use of these results in the statistical analysis of longitudinal data on premalignant lesions is described in Section 4. Potential applications and further developments are discussed in Section 5.

2. The number of premalignant clones

Following Dewanji et al. [7], we define a premalignant clone as the collection of all cells descended from a single progenitor intermediate cell that has risen from mutation of a normal cell. Note that because such cells can die or differentiate, a premalignant clone can regress or even become extinct. Suppose that observations on the same subject are taken at times $0 = t_0 < t_1 < \cdots < t_K < t_{K+1} = \infty$, where $K \geqslant 2$. Let N_{ii} denote the number of clones appearing in the interval $(t_{i-1}, t_i]$ and observed as non-extinct at time t_i, for $i = 1, \ldots, K$. In addition, suppose we also observe at time t_i, by some labelling technique, N_{li}, the number of clones appearing in $(t_{l-1}, t_l]$ and still observed (non-extinct) at time t_i, for $l = 1, \ldots, i-1$. Thus, the total number of premalignant clones observed at time t_i is given by $N_i = N_{1i} + \cdots + N_{ii}$, for $i = 1, \ldots, K$. Note that for fixed l, N_{li} is non-increasing in i because of possible extinction of one or more clones.

We transform the counts $\{N_{li}; l = 1, \ldots, i,\ i = 1, \ldots, K\}$ as follows. At any time t_i, we observe N_{ii} and subsequently M_{ij}, where M_{ij} denotes the number of clones appearing in $(t_{i-1}, t_i]$ and becoming extinct in $(t_j, t_{j+1}]$ for $j = i, \ldots, K$. Note that M_{iK} is the number of clones appearing in $(t_{i-1}, t_i]$ and possibly becoming extinct only after t_K, the last observation time. Note also that $N_{ii} = \sum_{j=i}^{K} M_{ij}$. Although we do not directly observe the M_{ij}s at time t_i, we define $C_i = \{N_{ii}, M_{ij}; j = 1, \ldots, K\}$ as the transformed event at time t_i. Note that the sets C_i are ordered naturally in time in the sense that C_i contain no observations made before time t_i. The C_is are also independent and collectively describe the whole of the count data. So it is enough to derive the probability of the C_is and then take the product over them to construct the likelihood function.

In order the derive the probability of C_i, note that N_{ii} follows a Poisson distribution with mean

$$\Lambda^{(i)}(t_i) = \int_{t_{i-1}}^{t_i} \lambda^{t_i}(s)\,\mathrm{d}s, \tag{1}$$

where

$$\lambda^{t_i}(s) = v(s)X(s)(1 - p(t_i, s)),$$

for $t_{i-1} < s \leqslant t_i$, where $p(t,s)$ denotes the probability that a clone appearing at time s becomes extinct by time t and is equal to $1-[g(t,s)+G(t,s)]^{-1}$ (see Eq. (10) in Section 3) with $g(t,s)=\exp\left[-\int_s^t\{\alpha(u)-\beta(u)\}\,\mathrm{d}u\right]$, and $G(t,s)=\int_s^t \alpha(u)g(u,s)\,\mathrm{d}u$ (see Refs. [7,16]). Given $N_{ii}=n_{ii}$, it is easily seen that $(M_{ii},\ldots,M_{iK})$ follows a multinomial distribution Multinomial$(n_{ii};p_{ii},\ldots,p_{iK})$ with parameters

$$p_{ij}=\frac{1}{\Lambda^{(i)}(t_i)}\int_{t_{i-1}}^{t_i} v(s)X(s)(p(t_{j+1},s)-p(t_j,s))\,\mathrm{d}s, \tag{2}$$

$(j=i,\ldots,K-1)$, and

$$p_{iK}=\frac{1}{\Lambda^{(i)}(t_i)}\int_{t_{i-1}}^{t_i} v(s)X(s)(1-p(t_K,s))\,\,\mathrm{d}s. \tag{3}$$

Consider now the case in which clones were not labelled at each observation time so that only the total counts N_i are available at time t_i. For $j>i$, note that N_{ij}, the number of clones appearing in $(t_{i-1},t_i]$ and still non-extinct at time t_j, is equal to $\sum_{l=j}^{K} M_{il}$. Hence, given $N_{ii}=n_{ii}$, N_{ij} follows a binomial distribution Binomial (n_{ii},q_{ij}), where

$$q_{ij}=\sum_{l=j}^{K}p_{il}=\frac{1}{\Lambda^{(i)}(t_i)}\int_{t_{i-1}}^{t_i} v(s)X(s)(1-p(t_j,s))\,\,\mathrm{d}s. \tag{4}$$

Hence, the unconditional distribution of N_{ij} is Poisson with mean $\Lambda^{(i)}(t_i)q_{ij}$. Since, for fixed j, the N_{ij}, for $i=1,\ldots,j$, are independent and $N_j=\sum_{i=1}^{j}N_{ij}$, N_j follows a Poisson distribution with mean

$$\left(\sum_{i=1}^{j}\Lambda^{(i)}(t_i)q_{ij}\right) \tag{5}$$

with $q_{jj}=1$, for $j=1,\ldots,K$. Note that

$$\sum_{i=1}^{j}\Lambda^{(i)}(t_i)q_{ij}=\int_0^{t_j} v(s)X(s)(1-p(t_j,s))\,\,\mathrm{d}s.$$

Although Eq. (5) is a known result (Ref. [7]), the approach used here to derive it will be useful for calculating covariances as will be seen later.

Since the N_i, for $i=1,\ldots,K$, are not independent, their joint distribution can be derived inductively. Consider first the simplest case of $K=2$. In this case, $N_1=N_{11}\sim$ Poisson$(\Lambda^{(1)}(t_1))$ and $N_2=N_{12}+N_{22}$ with $N_{22}\sim$ Poisson$(\Lambda^{(2)}(t_2))$. Since the conditional distribution of N_{12} given N_1 is Binomial(N_1,q_{12}), we have

$$\Pr\{N_2 = n_2 | N_1 = n_1\}$$
$$= \sum_{x=0}^{\min\{n_1,n_2\}} \frac{\mathrm{e}^{-\Lambda^{(2)}}(t_2)(\Lambda^{(2)}(t_2))^{n_2-x}}{(n_2-x)!} \binom{n_1}{x} q_{12}^{x}(1-q_{12})^{n_1-x}.$$

This approach can be extended for any $K > 2$, although the expressions quickly become so complex as to be impractical for application. However, the first two moments of the N_is can be obtained with less difficulty.

Specifically, for $i < j$, we have

$$\mathrm{cov}(N_i, N_j) = \mathrm{cov}\left(\sum_{l=1}^{i} N_{li}, \sum_{l=1}^{j} N_{lj}\right) = \sum_{i=1}^{i} \mathrm{cov}(N_{li}, N_{lj}) \tag{6}$$

since $\mathrm{cov}(N_{li}, N_{l'j}) = 0$ for $l \neq l'$. Now,

$$\mathrm{cov}(N_{li}, N_{lj}) = \mathrm{cov}\left(\sum_{u=i}^{K} M_{lu}, \sum_{v=j}^{K} M_{lv}\right)$$
$$= \sum_{u=i}^{K}\sum_{v=j}^{K} \mathrm{cov}(M_{lu}, M_{lv}). \tag{7}$$

Using Eq. (1) and the multinomial distribution in Eq. (2) and Eq. (3), we have

$$\mathrm{cov}(M_{lu}, M_{lv}) = E[-N_{ll}p_{lu}p_{lv}] + V[N_{ll}]p_{lu}p_{lv} = 0$$

for $u \neq v$. When $u = v$,

$$\mathrm{cov}(M_{lu}, M_{lv}) = V(M_{lu}) = E[N_{ll}p_{lu}(1 - p_{lu})] + V[N_{ll}p_{lu}] = p_{lu}\Lambda^{(l)}(t_l).$$

It follows, from Eq. (7) that

$$\mathrm{cov}(N_{li}, N_{lj}) = \sum_{u=j}^{K} p_{lu}\Lambda^{(l)}(t_l) = q_{lj}\Lambda^{(l)}(t_l),$$

and, from Eq. (6),

$$\mathrm{cov}(N_i, N_j) = \sum_{l=1}^{i} q_{lj}\Lambda^{(l)}(t_l) = \int_0^{t_i} v(s)X(s)(1 - p(t_j, s))\,\mathrm{d}s$$
$$= V_{ij}\ (\mathrm{say}). \tag{8}$$

3. The number and size of premalignant clones

Let $W_{ij}^{(l)}$ denote the size of the jth clone at time t_i arising from an intermediate cell generated during the interval $(t_{l-1}, t_l]$, for $j = 1, \ldots, N_{ll}$; $l = 1, \ldots, i$; $i = 1, \ldots, K$. Since premalignant clones may become extinct, it is possible that some of the $W_{ij}^{(l)}$s are zero. Clearly, if $W_{ij}^{(l)} = 0$ for some i, then $W_{i'j}^{(l)} = 0$ for all $i' > i$. For $l \leqslant i$, define $W_{ij}^{(l)}(s)$ as the size of the clone at time t_l

arising from an intermediate cell generated at time $S \in (t_{l-1}, t_l]$. Note that $W_{ij}^{(l)}(s)$ follows a birth–death process with birth and death rates $\alpha(\cdot)$ and $\beta(\cdot)$, respectively. Dewanji et al. [7,8] and Luebeck and Moolgavkar [16] show that

$$\Pr\{W_{ij}^{(l)}(s) = m\} = \frac{1}{G(t_i, s)} H(t_i, s)\{1 - H(t_i, s)\}^m, \tag{9}$$

for $m \geqslant 1$, and

$$\Pr\{W_{ij}^{(l)}(s) = 0\} = 1 - \frac{1}{g(t_i, s) + G(t_i, s)} = p(t_i, s), \tag{10}$$

where $g(t, s)$ and $G(t, s)$ are as defined in Section 2 following Eq. (1) and $H(t, s) = g(t, s)[g(t, s) + G(t, s)]^{-1}$. They also noted that the conditional distribution of $W_{ij}^{(l)}(s)$, given that the clone is non-extinct (that is, $W_{ij}^{(l)}(s) > 0$), is geometric with

$$\Pr\{W_{ij}^{(l)}(s) = m | W_{ij}^{(l)}(s) > 0\} = H(t_i, s)\{1 - H(t_i, s)\}^{m-1}, \tag{11}$$

from $m \geqslant 1$.

Let E_i denote the event $\{N_{ii}, W_{ij}^{(l)}; j = 1, \ldots, N_{ll},\ l = 1, \ldots, i\}$ at time t_i. Note that the sequence of events $\{E_i\}_{i=1}^K$ forms a Markov process. Hence, the probability of the observation from a single subject $\Pr\{E_1, \ldots, E_K\}$ is derived by 'successive conditioning' as follows.

For the first event E_1 at time t_1, N_{11} follows a Poisson distribution with mean $\Lambda^{(1)}(t_1)$, as in Eq. (1). Given $N_{11} = n_{11}$, the clone sizes $W_{1j}^{(1)}$, for $j = 1, \cdots, n_{11}$, are independent and identically distributed with

$$\Pr\{W_{1j}^{(1)} = m\} = \frac{1}{\Lambda^{(1)}(t_1)} \int_0^{t_1} \lambda^{t_1}(s) H(t_1, s)(1 - H(t_1, s))^{m-1}\, \mathrm{d}s, \tag{12}$$

for $m \geqslant 1$ (see Ref. [7]). The probability of the event E_1 is now given by $\Pr\{N_{11} = n_{11}\}$ multiplied by a product of n_{11} terms.

In general, for $i > 1$, we want to find $\Pr\{E_i | E_1, \ldots, E_{i-1}\} = \Pr\{E_i | E_{i-1}\}$. Note that N_{ii} follows, as in Eq. (1), a Poisson distribution with mean $\Lambda^{(i)}(t_i)$. As in Eq. (12), given $N_{ii} = n_{ii}$, the $W_{ij}^{(i)}$, for $j = 1, \ldots, n_{ii}$ are independent and identically distributed with

$$\Pr\{W_{ij}^{(i)} = m\} = \frac{1}{\Lambda^{(i)}(t_i)} \int_{t_{i-1}}^{t_i} \lambda^{t_i}(s) H(t_i, s)(1 - H(t_i, s))^{m-1}\, \mathrm{d}s, \tag{13}$$

for $m \geqslant 1$. Note that this part of E_i, namely N_{ii} and $W_{ij}^{(i)}$, for $j = 1, \ldots, N_{ii}$, is independent of E_{i-1}. However, for $l = 1, \ldots, i-1$, the sizes $W_{ij}^{(l)}$, for $j = 1, \ldots, N_{ll}$, at time t_i depend on the corresponding sizes $W_{i-1,j}^{(l)}$ at time t_{i-1}, which is a part of E_{i-1}. Given E_{i-1}, note that $W_{ij}^{(l)}$ is a birth–death process with birth and death rates $\alpha(\cdot)$ and $\beta(\cdot)$, respectively, starting at time t_{i-1} with initial size $W_{i-1,j}^{(l)} = w_{i-1,j}^{(l)}$. The probability generating function of $W_{ij}^{(l)}$, given E_{i-1}, at time t_i is then

$$\begin{aligned}\Phi(x;t_i,t_{i-1}) &= \left[1 - \frac{x-1}{(x-1)G(t_i,t_{i-1}) - g(t_i,t_{i-1})}\right]^{w_{i-1,j}^{(l)}} \\ &= \left[\Psi(x;t_i,t_{i-1})\right]^{w_{i-1,j}^{(l)}}, \text{ say,} \end{aligned} \tag{14}$$

where $\Psi(x;t_i,t_{i-1})$ is the probability generating function of a birth–death process with rates $\alpha(\cdot)$ and $\beta(\cdot)$, respectively, starting at time t_{i-1} with initial size 1 (see Ref. [8]), leading to the probability mass function $p(\cdot)$ of the form (9) and (10). It is now routine, at least in theory, from Eq. (14), to obtain the conditional probability mass function of $W_{ij}^{(l)}$ given E_{i-1}, for $l = 1, \ldots, i-1$.

In particular, for $i \geqslant 2, l = 1, \ldots, i-1$, note that, given $W_{i-1,j}^{(l)} = w_{i-1,j}^{(l)}$ which is a part of $E_{i-1}, W_{ij}^{(l)}$ can be written as $\sum_{u=1}^{w_{i-1,j}^{(l)}} W_u$, where the W_us are independent and identically distributed with probability generating function $\Psi(x;t_i,t_{i-1})$. Then, we have Eq. (1)

$$\begin{aligned}\Pr\{W_{ij}^{(l)} = m\} &= \sum_{m_1,\ldots m_{w_{i-1,j}^{(l)}} : \sum m_u = m} \prod_{u=1}^{w_{i-1,j}^{(l)}} p(m_u) = (1-H)^m \sum_{v=0}^{w_{i-1,j}^{(l)}} \binom{w_{i-1,j}^{(l)}}{v} \\ &\quad \times c(m, w_{i-1,j}^{(l)} - v) \times \left(1 - \frac{1}{g+G}\right)^v \left(\frac{1}{G}H\right)^{w_{i-1,j}^{(l)} - v}\end{aligned} \tag{15}$$

for $m \geqslant 0$, where the arguments of g, G and H are t_i and t_{i-1}, respectively, and

$$c(m,n) = \#\left\{(m_1,\ldots,m_n):\ m_i > 0, \text{ for all } i, \sum_{i=1}^{n} = m\right\} = \binom{m-1}{n-1} \tag{16}$$

for $m \geqslant n$ with $c(0,0) = 1$, and 0 for $m < n$. Putting Eq. (16) into Eq. (15) gives the probability mass function of $W_{ij}^{(l)}$ as a sum of $(w_{i-1,j}^{(l)} + 1)$ terms. For large $w_{i-1,j}^{(l)}$, one can try some approximation for this sum. However, the probability $\Pr\{E_i|E_{i-1}\}$ can be calculated by multiplying $\Pr\{N_{ii} = n_{ii}\}$ (see (1)) by a product of n_{ii} terms like Eq. (13) and that of $\sum_{l=1}^{i-1} n_{li}$ probability terms like Eq. (15).

4. Statistical considerations

The distributional results given in Sections 2 and 3 provide a statistical basis for fitting the two stage model to experimental data. When the distribution of the observations is completely specified, estimates of the unknown model parameters may be obtained by maximizing the likelihood of the data. Suppose first that full information on the number and sizes of premalignant clones is available through $E_1, \ldots, E_K$ (Section 3). The contribution to the likelihood by a single subject is then proportional to $\Pr\{E_1\} \times \Pr\{E_2|E_1\} \times \cdots \times \Pr\{E_K|E_{K-1}\}$. The full likelihood is proportional to the product of the likelihood contributions from all experimental subjects.

The likelihood function for the case, in which only the number of clones, but not their sizes, is observed (Section 2) along with some labelling so as to identify them regarding which time interval they appeared in and became extinct, if so, can be constructed using $\Pr\{N_{ii} = n_{ii}\}$ from Eq. (1) and the multinomial structure (2) and (3). In the absence of such labelling, when at each time point only the total number of clones is observed, the likelihood function can be written down in theory, although the expressions are unmanageable for $K > 2$. To circumvent this problem, model fitting methods that require specification only of the first two moments of the data may be employed.

Specifically Eq. (5) and Eq. (8) can now be used as the basis for an iteratively reweighted least squares method in which the sum of squares

$$SS = \sum (\mathbf{N} - \boldsymbol{\mu})^{\mathrm{T}} V^{-1} (\mathbf{N} - \boldsymbol{\mu})$$

is minimized with respect to the model parameters, where the sum is over all subjects. Here $\mathbf{N}^{\mathrm{T}} = (N_1, \ldots, N_k)$, and $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_1, \ldots, \mu_k)$ with

$$\mu_i = \sum_{l=1}^{i} q_{li} \Lambda^{(l)}(t_l) = \int_0^{t_i} v(s) X(s)(1 - p(t_i, s))\,\mathrm{d}s,$$

and V_{ij}, the (i,j)th entry of V, the variance (weight) matrix, is defined in Eq. (8) for $i < j$ with $V_{ji} = V_{ij}$ and $V_{ii} = \mu_i$.

In fitting the two stage model to experimental data, it is necessary to specify the manner in which the model parameters vary over time. In typical application, the first stage mutation rate vX and the birth and death rates α and β of intermediate cells are allowed to vary as functions of dose (see Refs. [19,21]). In the simplest case in which the dose is held constant over time, the expressions derived in Sections 2 and 3 reduce to simple forms. For example, Dewanji et al. [7] show that

$$p(t,s) = \frac{\beta - \beta \exp - (\alpha - \beta)(t - s)}{\alpha - \beta \exp - (\alpha - \beta)(t - s)}. \tag{17}$$

It follows that

$$\begin{aligned} \Lambda^{(i)}(t_i) &= vX \int_{t_{i-1}}^{t_i} \frac{\alpha - \beta}{\alpha - \beta \exp - (\alpha - \beta)(t_i - s)}\,\mathrm{d}s \\ &= \frac{vX}{\alpha} \left[\log \left(\frac{\alpha \exp (\alpha - \beta)(t_i - t_{i-1}) - \beta}{\alpha - \beta} \right) \right], \end{aligned} \tag{18}$$

and

$$
\begin{aligned}
q_{ij}\Lambda^{(i)}(t_i) &= vX\int_{t_{i-1}}^{t_i} \frac{\alpha-\beta}{\alpha-\beta\exp-(\alpha-\beta)(t_j-s)}\,\mathrm{d}s \\
&= \frac{vX}{\alpha}\left[\log\left(\frac{\alpha\exp(\alpha-\beta)(t_j-t_{i-1})-\beta}{\alpha\exp(\alpha-\beta)(t_j-t_i)-\beta}\right)\right]. \qquad (19)
\end{aligned}
$$

Noting that $p_{iK} = q_{iK}$ and $p_{ij} = q_{ij} - q_{i,j+1}$, for $j = i,\ldots,K-1$, we obtain simple forms of the expressions necessary for analyzing count data.

Simplification of the expression needed in the analysis of data on clone sizes also occurs in the case of time-homogeneous model parameters. Following Moolgavkar et al. [19], Eq. (13) reduces to

$$
\Pr\left\{W_{ij}^{(i)} = m\right\} = \frac{\left[\frac{\alpha}{\beta}p(t_i-t_{i-1})\right]^m}{m\log\left(\frac{\beta}{\beta-\alpha\ p(t_i-t_{i-1})}\right)}, \qquad (20)
$$

for $m \geqslant 1$. The preceding results apply in the case $\alpha \neq \beta$. Similar results for the case $\alpha = \beta$ are readily obtained, but are omitted for brevity.

Similar results can be obtained when the parameter values are not constant over time, but remain constant within specified time intervals. Computable forms of Eqs. (17)–(20) can also be obtained in this case, although the expressions are somewhat lengthy.

In deriving the distributional results given in Sections 2 and 3, it was tacitly assumed that a (non-extinct) premalignant clone remains visible regardless of its size. In reality, a clone may not be detectable unless it exceeds a certain threshold, involving a minimum of $n_0 > 0$ intermediate cells. If a threshold for the identification of premalignant clones is considered, the distribution of the observed number of clones N_i at time t_i is again Poisson, but with the function $p(t;s)$ in the expression for the mean of this distribution replaced by $p^*(t;s)$, the probability of non–detection Ref. [7]. Other distributional results are more complex, because non-detection, unlike extinction, is not an absorbing state. Since a clone can oscillate between detectable and non-detectable states, the multinomial structure in Eq. (2) and Eq. (3) no longer holds.

Since this oscillation occurs with low probability, it is reasonable to assume, as a first approximation, that after a once-detectable clone has become non-detectable, it does not return back to a detectable state again. Under this assumption, the multinomial structure in Eq. (2) and Eq. (3) holds, with cell probabilities p_{ij} reflecting the probability of a clone being non-detectable in $(t_j, t_{j+1}]$ given that it was detected in $(t_{i-1}, t_i]$ for $j = i,\ldots,K$. It can then be shown that the covariance function $\mathrm{cov}(N_i,\ N_j)$ for detectable clones is approximately of the form (8), but with $p(t_j, s)$ replaced by $p^*(t_j, s)$. With these approximate covariances, generalized estimating equations (GEEs) can be used for model fitting, with parameter estimates obtained by solving the estimating equations

$$\sum \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} V^{-1}(\mathbf{N} - \boldsymbol{\mu}) = \mathbf{0}$$

(see Ref. [15]) for $\boldsymbol{\theta}$, the vector of parameters of interest, where V now denotes the approximate variance matrix. GEEs have been used in fitting dose response models to data from experiments on mutagenicity (Ref. [14]) and developmental toxicity (Ref. [26]).

In initiation-promotion experiments involving the mouse skin system, there can be appreciable variation in response among individual animals, as noticed by Burnett et al. [3]. They propose the use of non-linear random-effects regression modelling techniques to describe such inter-individual variation. These techniques can be applied in fitting the two stage model to experimental data involving multiple subjects by allowing for variations in the model parameters among individuals.

5. Concluding remarks

In this article, we have discussed the use of the two stage clonal expansion model of carcinogenesis to describe the number and size of premalignant clones in experiments in which such clones are observable. A novel feature of our analysis is the provision for repeated observations on the same individual over time. Assuming that even very small clones are detectable, the joint distribution of the number and size of clones observed on a single subject at different points in time can be derived. Approximate distributional results can also be obtained when clones can be observed only when they achieve a specified size.

These distributional results (Sections 2 and 3) provide a basis for fitting the two stage model to longitudinal data on premalignant clones derived from laboratory experiments. When the joint distribution of the observations is fully specified, estimates of the model parameters can be obtained by the method of maximum likelihood. When only the first two moments of the observations are specified, iteratively weighted least squares method or generalized estimating equations can be used for parameter estimation. Random effects modelling techniques can be used to allow for heterogeneity in the values of the model parameters among individuals.

Kopp-Schneider and Portier [12] consider a modification of the two stage model in which, at iniation, one of two types of initiated cells is produced, one type is terminally benign or premalignant and never progresses towards carcinoma or malignancy, and the other may progress to become malignant. This inclusion of two pathways may explain some observed heterogeneity in the population of papillomas (see also Ref. [4]). The distributional results under this modified model will be more difficult; however, some efforts in this direction are under way. Extensions of the results presented in this paper to permit a joint analysis of data on premalignant and malignant lesions may be possible. However, because of the presence of serial correlation in longitudinal data of the type considered here, this analysis will be more complicated than previous analysis of cross-sectional data given by Dewanji et al. [8] and de Gunst and Luebeck [6].

Acknowledgements

This research was supported by an American Cancer Society International Cancer Research Fellowship from UICC to A.D., and by an NCI grant 47658 to S.H.M.

References

[1] P. Armitage, R. Doll, The age distribution of cancer and a multistage theory of carcinogenesis, Br. J. Cancer 8 (1954) 1.

[2] R.K. Boutwell, Model systems for defining initiation, promotion, and progression of skin neoplasms, Prog. Clinical Biol. Res. 298 (1989) 3.

[3] R.T. Burnett, W.H. Ross, D. Krewski, Nonlinear random effects regression models, Environmetrics 6 (1995) 85.

[4] K.C. Chu, C.C. Brown, R.E. Tarone, W.Y. Tan, Differentiating among proposed mechanisms for tumor promotion in mouse skin with the use of the multievent model for cancer, J. Nat. Cancer Inst. 79 (1987) 789.

[5] J. Denes, D. Krewski, An exact representation of the generating function for the Moolgavkar–Venzon–Knudson two stage model of carcinogenesis with stochastic stem cell growth, Math. Biosci. 131 (1996) 185.

[6] M. de Gunst, E.G. Luebeck, Quantitative analysis of two-dimensional observation of premalignant clones in the presence or absence of malignant tumors, Math. Biosci. 119 (1994) 5.

[7] A. Dewanji, D.J. Venzon, S.H. Moolgavkar, A stochastic two-stage model for cancer risk assessment. II. The number and size of premalignant clones, Risk Anal. 9 (1989) 179.

[8] A. Dewanji, S.H. Moolgavkar, E.G. Luebeck, Two mutation model for carcinogenesis: Joint analysis of premalignant and maligant lesions, Math. Biosci. 104 (1991) 97.

[9] D.G. Kendall, Birth-and-death processes, and the theory of carcinogenesis, Biometrika 47 (1960) 13.

[10] A. Kopp-Schneider, C.J. Portier, Birth and death/differentiation rates of papillomas in mouse skin, Carcinogenesis 12 (1992) 973.

[11] A. Kopp-Schneider, C.J. Portier, A stem cell model for carcinogenesis, Math. Biosci. 120 (1994) 211.

[12] A. Kopp-Schneider, C.J. Portier, Carcinoma formation in mouse skin painting studies is a process suggesting greater than two stage, Carcinogenesis 16 (1995) 53.

[13] D. Krewski, M. Goddard, J. Zielinski, Dose response relationships in carcinogenesis, in: H. Vainio, P. Magee, D. McGregor, A. McMichael (Eds.), Mechanisms of Carcinogenesis in Risk Identification, International Agency for Research on Cancer, Lyon, 1992.

[14] D. Krewski, B. Leroux, S. Bleuer, L. Broekhoven, Modelling the Ames Salmonella/microsome Assay, Biometrics 49 (1993) 499.

[15] K.-Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13.

[16] E.G. Luebeck, S.H. Moolgavkar, Stochastic analysis of intermediate lesions in carcinogenesis experiments, Risk Anal. 11 (1991) 149.

[17] J. McLean, M. Stuchly, R. Mitchel, D. Wilkinson, H. Yang, M. Goddard, D. Lecuyer, M. Schunk, E. Callary, D. Morrison, Cancer promotion in a mouse-skin model by a 60-hz magnetic fields: I. Tumor development and immune response, Bioelectro Magn. 12 (1991) 273.

[18] S. Moolgavkar, A. Knudson, Mutation and cancer: A model for human carcinogenesis, J. Nat. Cancer Inst. 66 (1981) 1037.

[19] S.H. Moolgavkar, E.G. Luebeck, M. de Gunst, R. Port, M. Schwartz, Quantitative analysis of enzyme-altered foci in rat hepatocarcinogenesis experiments, Carcinogenesis 11 (1990) 1271.

[20] S.H. Moolgavkar, E.G. Luebeck, Two-event model for carcinogenesis: Biological, mathematical and statistical considerations, Risk Anal. 10 (1990) 323.

[21] S.H. Moolgavkar, E.G. Luebeck, D. Krewski, J. Zielinski, Randon, cigarette smoke, and lung cancer: A re-analysis of the Colorado plateau uranium miners' data, Epidemiology 4 (1993) 194.

[22] S.H. Moolgavkar, D.J. Venzon, Two-event models for carcinogenesis: Incidence curves for childhood and adult tumors, Math. Biosci. 47 (1979) 55.

[23] C. Portier, A. Kopp-Schneider, A multistage model for carcinogenesis incorporating damage and repair, Risk Anal. 11 (1991) 535.

[24] M. Schwarz, D. Pearson, A. Buchman, W. Kunz, The use of enzyme-altered foci for risk assessment of hepatocarcinogenesis. in: C. Travis (Ed.), Biologically Based Methods for Cancer Risk Assessment, NATO ASI Series A: Life Sciences, vol. 159, Plenum, New York, 1989, 31.

[25] W.Y. Tan, Stochastic Models of Carcinogenesis. Marcel Dekker, New York, 1991.

[26] Y. Zhu, D. Krewski, W.H. Ross, Dose response models for overdispersed multinomial data from development toxicity experiments, Appl. Statist. 43 (1994) 583.