Multistage clonal expansion model hazard derivation

The derivation of the multistage clonal expansion model hazards has been previously shown (e.g. in the appendices of Luebeck and Moolgavkar (2002) and Jeon (2007).) We reproduce it here with our notation.

The \( n \)-stage clonal expansion model is a continuous-time Markov chain with the following states:

- \( \mathbf{X}(t) \), the number of normal cells at age \( t \);
- \( \mathbf{Y}_1(t), \ldots, \mathbf{Y}_{n-2}(t) \), the number of cells in subsequent preinitiation states;
- \( \mathbf{Y}_{n-1}(t) \), the number of initiated cells; and
- \( Z(t) \), the number of malignant cells.

Let \( \nu \) be the initial mutation rate (with \( \mu_0 := \nu X \)), \( \mu_1, \ldots, \mu_{n-3} \) the following preinitiation mutation rates, \( \mu_{n-2} \) the initiation mutation rate, \( \mu_{n-1} \) the malignant transformation rate, \( \alpha \) the clonal expansion rate, and \( \beta \) the cell death rate. Mutations are modeled as an asymmetric division process resulting in one cell of the same class as the parent and one of the subsequent class. For notational simplicity, we suppress the possible dependence of the parameters on time.

Let

\[
P_{(j_1, \ldots, j_{n-1}, k)}(t) = P[\mathbf{Y}_1(t) = j_1, \ldots, \mathbf{Y}_{n-1}(t) = j_{n-1}, Z(t) = k | \mathbf{Y}_1(0) = 0, \ldots, \mathbf{Y}_{n-1}(0) = 0, Z(0) = 0].
\]

(S1)

Then, the probability generating function for the numbers of preinitiated, initiated, and malignant cells may be written as

\[
\Psi(y_1, \ldots, y_{n-1}, z, t) = E \left[ y_1^{Y_1(t)} \cdots y_{n-1}^{Y_{n-1}(t)} z^Z(t) | \mathbf{Y}_1(0) = 0, \ldots, \mathbf{Y}_{n-1}(0) = 0, Z(0) = 0 \right],
\]

(S2)

\[
= \sum_{(j_1, \ldots, j_{n-1}, k)} P_{(j_1, \ldots, j_{n-1}, k)}(t)y_1^{j_1} \cdots y_{n-1}^{j_{n-1}} z^k.
\]

(S3)

We wish to solve for the survival and hazard functions,

\[
S_n(t) = \sum_{(j_1, \ldots, j_{n-1}, 0)} P_{(j_1, \ldots, j_{n-1}, 0)}(t),
\]

(S4)

\[
= \sum_{(j_1, \ldots, j_{n-1}, k)} P_{(j_1, \ldots, j_{n-1}, k)}(t) 1^{j_1} \cdots 1^{j_{n-1}} 0^k,
\]

(S5)

\[
= \Psi(1, \ldots, 1, 0, t),
\]

(S6)

\[
h_n(t) = -\frac{d}{dt} \ln \Psi \bigg|_{(1, \ldots, 1, 0, t)}.
\]

(S7)

To this end, we write the Kolmogorov forward equations for the transition probabilities of this
process,
\[
\frac{d}{dt} P_{(j_1, \ldots, j_{n-1}, k)} = \mu_0 P_{(j_1-1, \ldots, j_{n-1}, k)} + j_1 \mu_1 P_{(j_1, j_2-1, \ldots, j_{n-1}, j_1-1, k)} + \cdots + j_{n-2} \mu_{n-2} P_{(j_1, \ldots, j_{n-2}, j_{n-1}+1, k)} + j_{n-1} \mu_{n-1} P_{(j_1, \ldots, j_{n-1}, k-1)} + \alpha(j_{n-1} - 1) P_{(j_1, \ldots, j_{n-1}-1, k)} + \beta(j_{n-1} + 1) P_{(j_1, \ldots, j_{n-1}+1, k)} - (\mu_0 + j_1 \mu_1 + \cdots + j_{n-2} \mu_{n-2} + j_{n-1}(\alpha + \beta + \mu_{n-1})) P_{(j_1, \ldots, j_{n-1}, k)}.
\]
(S8)

Multiplying both sides by \(y_1^{j_1} \cdots y_{n-1}^{j_{n-1}} z^k\) and summing over all \((j_1, \ldots, j_{n-1}, k)\), we arrive at the forward Kolmogorov equation for the probability generating function,
\[
\frac{\partial \Psi}{\partial t} = \left( y_1 - 1 \right) \mu_0 \Psi + \left( y_2 - 1 \right) \mu_1 y_1 \frac{\partial \Psi}{\partial y_1} + \cdots + \left( y_{n-1} - 1 \right) \mu_{n-2} y_{n-2} \frac{\partial \Psi}{\partial y_{n-1}} + \left( \alpha y_{n-1}^2 + \beta - (\alpha + \beta + \mu_{n-1}(1 - z)) y_{n-1} \right) \frac{\partial \Psi}{\partial y_{n-1}}.
\]
(S9)

Because all cells are assumed to be normal at time \(t = 0, \Psi = 1\) at this time. The partial differential equation in Eq. (S9) is first-order and thus may be solved by the method of characteristics. We choose characteristic curves \((y_1(\tau), \ldots, y_{n-1}(\tau), z(\tau), t(\tau))\) such that the following ordinary differential equations are satisfied:
\[
\begin{align*}
\frac{dy_1}{d\tau} &= (1 - y_2) \mu_1 y_1, \\
\vdots & \\
\frac{dy_{n-2}}{d\tau} &= (1 - y_{n-1}) \mu_{n-2} y_{n-2}, \\
\frac{dy_{n-1}}{d\tau} &= (\alpha + \beta + \mu_{n-1}(1 - z)) y_{n-1} - \alpha y_{n-1}^2 - \beta, \\
\frac{dz}{d\tau} &= 0, \\
\frac{dt}{d\tau} &= 1.
\end{align*}
\]
(S10)

We then write Eq. (S9) as an ordinary differential equation,
\[
\frac{d\Psi}{d\tau} = (y_1(\tau) - 1) \mu_0 \Psi,
\]
(S11)
which has the solution
\[
\Psi(\tau) = \exp \left[ \int_{\tau_0}^{\tau} (y_1(u) - 1) \mu_0 \, du \right],
\]
(S12)
where \(\tau_0\) is such that \(t(\tau_0) = 0\), so that \(\Psi(\tau_0) = 1\) by our initial condition.
Because this equation depends on $y_1$, we must further consider the system of equations in Eqs. (S10). Because we are solving for $S(t) = \Psi(1, \ldots, 1, 0, t)$, we consider the characteristic curve, that is $(y_1(\tau), \ldots, y_{n-1}(\tau), z(\tau), t(\tau))$, with boundary condition $(1, \ldots, 1, 0, t^*)$, where $t^*$ denotes a fixed $t$. Along this characteristic, $t = \tau - \tau_0$, $z = 0$, and $y_{n-1}$ satisfies a Ricatti equation (Heidenreich et al., 1997), which, in the case of constant parameters, has the solution (with $\tau^* = \tau_0 + t^*$)

$$y_{n-1}(\tau) = \frac{p_nq_n}{\alpha} \left( \frac{e^{-p_n(\tau^*-\tau)} - e^{-q_n(\tau^*-\tau)}}{q_ne^{-p_n(\tau^*-\tau)} - p_ne^{-q_n(\tau^*-\tau)}} \right) + 1, \quad (S13)$$

where

$$p_n, q_n := \frac{1}{2} \left( - (\alpha - \beta - \mu_{n-1}) \pm \sqrt{(\alpha - \beta - \mu_{n-1})^2 + 4\alpha\mu_{n-1}} \right). \quad (S14)$$

Moreover, each $y_i(\tau)$, for $1 \leq i \leq n - 2$, can be found recursively, noting $y_i(\tau^*) = 1$,

$$y_i(\tau) = \exp \left[ \int_{\tau^*}^{\tau} \left(1 - y_{i+1}(u)\right)\mu_i \, du \right]. \quad (S15)$$

Although the backward formulation (see, for example, Brouwer et al. (2016)) is more computationally tractable, this forward formulation allows us to find formulas for the hazards of the first several models in the constant parameter case.

**Two-stage model**

For the two-stage model, we can analytically derive the solution for $\Psi$ in the constant parameter case.

$$\Psi(\tau) = \exp \left[ \int_{\tau_0}^{\tau} (y_1(u) - 1)\mu_0 \, du \right], \quad (S16)$$

$$= \exp \left[ \int_{\tau_0}^{\tau} \left( \frac{p_2q_2}{\alpha} \left( \frac{e^{-p_2(\tau^*-u)} - e^{-q_2(\tau^*-u)}}{q_2e^{-p_2(\tau^*-u)} - p_2e^{-q_2(\tau^*-u)}} \right) \right) \mu_0 \, du \right], \quad (S17)$$

$$= \left( \frac{q_2e^{-p_2(\tau^*-\tau)} - p_2e^{-q_2(\tau^*-\tau)}}{q_2e^{-p_2(\tau^*-\tau_0)} - p_2e^{-q_2(\tau^*-\tau_0)}} \right)^{\mu_0/\alpha}. \quad (S18)$$

Expressed as a function of $t$,

$$\Psi(t) = \left( \frac{q_2e^{-p_2(t^*-t)} - p_2e^{-q_2(t^*-t)}}{q_2e^{-p_2t^*} - p_2e^{-q_2t^*}} \right)^{\mu_0/\alpha}. \quad (S19)$$

Since $t^*$ is the value of $t$ such that $(y_1, \ldots, y_{n-1}, z) = (1, \ldots, 1, 0)$, $S(t^*) = \Psi(t^*)$. Viewing $S$ as a function of $t = t^*$,

$$S_2(t) = \left( \frac{q_2 - p_2}{q_2e^{-p_2t} - p_2e^{-q_2t}} \right)^{\mu_0/\alpha}, \quad (S20)$$

and

$$h_2(t) = \frac{\mu_0}{\alpha} \left( \frac{p_2q_2 (e^{-q_2t} - e^{-p_2t})}{q_2e^{-p_2t} - p_2e^{-q_2t}} \right). \quad (S21)$$
Three-stage model

For the three-stage model, we have

\[ \Psi(\tau) = \exp \left[ \int_{\tau_0}^{\tau} (y_1(u_1) - 1) \mu_0 du_1 \right], \]  
\( (S22) \)

\[ \Psi(\tau) = \exp \left[ \int_{\tau_0}^{\tau} \left( \exp \left[ \int_{\tau_0}^{u_1} (1 - y_2(u_2)) \mu_1 du_2 \right] - 1 \right) \mu_0 du_1 \right], \]  
\( (S23) \)

\[ \Psi(\tau) = \exp \left[ \int_{\tau_0}^{\tau} \left( \exp \left[ \int_{\tau_0}^{u_1} \left( - \frac{p_3 q_3}{\alpha} \left( e^{-p_4(\tau^*-u_2)} - e^{-q_4(\tau^*-u_2)} \right) \right) \mu_1 du_2 \right] - 1 \right) \mu_0 du_1 \right], \]  
\( (S24) \)

\[ \Psi(\tau) = \exp \left[ \int_{\tau_0}^{\tau} \left( \frac{q_3 - p_3}{q_3 e^{-p_3(\tau^*-u_1)} - p_3 e^{-q_3(\tau^*-u_1)}} \right)^{\mu_1/\alpha} \right] \mu_0 du_1 \right], \]  
\( (S25) \)

and

\[ S_3(t) = \exp \left[ \int_0^t \left( \frac{q_3 - p_3}{(q_3 e^{-p_3(t-u)} - p_3 e^{-q_3(t-u)})} \right)^{\mu_1/\alpha} \right] \mu_0 du \right]. \]  
\( (S26) \)

Although this survival function cannot be written in closed form, the hazard can be,

\[ h_3 = \frac{-d}{dt} \ln S_3(t), \]  
\( (S27) \)

\[ h_3 = \mu_0 \left( 1 - \left( \frac{q_3 - p_3}{q_3 e^{-p_3(t-u)} - p_3 e^{-q_3(t-u)}} \right)^{\mu_1/\alpha} \right). \]  
\( (S28) \)

Four-stage model

For the four-stage model, we continue to describe \( \Psi \) recursively,

\[ \Psi(\tau) = \exp \left[ \int_{\tau_0}^{\tau} \left( \exp \left[ \int_{\tau_0}^{u_1} \left( 1 - \left( \exp \left[ \int_{\tau_0}^{u_2} (1 - y_3(u_3)) \mu_2 du_3 \right] \right) \mu_1 du_2 \right] - 1 \right) \mu_0 du_1 \right] \right], \]  
\( (S29) \)

which leads to the following integral formula for the hazard

\[ h_4(t) = \nu X \left( 1 - \exp \left( \int_0^t \mu_1 \left( \frac{q_4 - p_4}{(q_4 e^{-p_4(t-u)} - p_4 e^{-q_4(t-u)})} \right)^{\mu_2/\alpha} \right) du \right). \]  
\( (S30) \)
Model likelihood

We assume that incidence of cancer is Poisson-distributed with mean number of cases

\[ \mu_i = N_i \cdot h_n(\rho_n, t_i) \]  

(S31)

at age \( t_i \) with population \( N_i \); here, \( n \in \{2, 3, 4\} \), and \( \rho_n \) is \( \{p_2, q_2, \nu X/\alpha\}, \{p_3, q_3, \nu X, \mu_1/\alpha\}, \) and \( \{p_4, q_4, \nu X, \mu_1, \mu_2/\alpha\} \), respectively. The likelihood can then be written as

\[ L(\rho_n) = \prod_i \frac{e^{-\mu_i} \mu_i^{x_i}}{x_i!} \]  

(S32)

where \( \{x_i\} \) are the numbers of observed cases. Maximizing the likelihood function—as a function of the parameters comprising \( \rho_n \)—is equivalent to minimizing the negative log-likelihood (NLL)

\[ \text{NLL}(\rho_n) = - \sum_i (\mu_i - x_i \log \mu_i - \log x_i!) \]  

(S33)