A new look at multi-stage models of cancer incidence

Tyler Lian* and Rick Durrett†
Dept. of Math, Duke U.,
P.O. Box 90320, Durham, NC 27708-0320

January 5, 2018

Abstract

Multi-stage models have long been used in combination with SEER data to make inferences about the mechanisms underlying cancer initiation. The main method for studying these models mathematically has been the computation of generating functions by solving hyperbolic partial differential equations. Here, we analyze these models using a probabilistic approach similar to the one Durrett and Moseley [7] used to study branching process models of cancer. This more intuitive approach leads to simpler formulas and new insights into the behavior of these models. Unfortunately, the examples we consider suggest that fitting multi-stage models has very little power to make inferences about the number of stages unless parameters are constrained to take on realistic values.

1 Introduction

Investigation of the age distribution of cancer incidence goes back to the middle of the 20th century. Fisher and Holloman [10] and Nordling [20] found that within the 25–74 age range, the logarithm of the cancer death rate increased in direct proportion to the logarithm of the age, with a slope of about six on a log-log plot. Nordling grouped all types of cancer together and considered only men, but the pattern persisted when Armitage and Doll [2] separated cancers by their type and considered men and women separately.

Nordling [20] suggested that the slope of six on a log-log plot would be explained if a cancer cell was the end result of seven successive mutations. There was no model underlying that conclusion, just the observation that if one sums k exponential random variables with rate μ_i, i.e., with probability density $\mu_i e^{-\mu_i t}$, then when t is much smaller than the mean of the sum $\sum_{i=1}^{k} 1/\mu_i$, then "the probability that the kth change occurs in the short time interval $(t, t+dt)$ is asymptotically

$$\frac{\mu_1, \mu_2 \cdots \mu_k t^{k-1}}{(k-1)!}\, dt." \tag{1}$$

*This research was done while TL was participating in the Huang Fellows Summer Program at Duke
†RD is partially supported by NSF grant DMS 1614838 from the math biology program.

Later Armitage [1] gave a rigorous proof of this result.

A few years later, Armitage and Doll [3] wrote that the hypothesis described in the previous paragraph was "however, unsatisfactory in that there was no direct experimental evidence to suggest that carcinogenesis was likely to involve more than two stages." Because of this, they introduced in [3] a two-stage model in which ordinary cells (type 0) mutate at rate μ_1 into initiated (type 1) cells that grow at exponential rate λ and mutate at rate μ_2 into malignant cells (type 2). The two-stage model has been thoroughly analyzed in the literature, see e.g. [12], [13]

In the studies cited above, the stages were unspecified events. That changed in 1971 with Knudson's study of retinoblastoma [14]. Based on observations of 48 cases of retinoblastoma and published reports, he hypothesized that the disease is a cancer caused by two mutational events. In the dominantly inherited form, one mutation is inherited via the germinal cells. In the nonhereditary form both mutations occur in somatic cells. The underlying gene, named RB1, was found 15 years later. In currentl terminology, RB1 is a *tumor suppressor gene*. Trouble begins when both copies are knocked out.

Colorectal tumors provide an excellent system in which to search for and study the genetic alterations involved in the development of cancer because tumors of various stages of development, from very small adenomas to very large carcinomas, can be obtained for study. The initiating event is thought to involve the inactivation of the tumor suppressor gene APC (adenomatous polyposis coli). As in retinoblastoma, an inherited germ line mutation in this gene causes greater risk of disease. Individuals with this mutation have numerous polyps form early in their lives, mainly in the epithelium of the large intestine.

In 1990, Fearon and Vogelstein [8] found a second piece of the puzzle when they noted that approximately 50% of colorectal carcinomas, and a similar percentage of adenomas greater than 1 cm have mutations in the RAS gene family, while only 10% of adenomas smaller than 1 cm have these mutations. In the modern terminology, the members of the RAS family are *oncogenes*. A mutation to a single allele is sufficient for progression. The analysis in [8] also suggested a role for TP53 (which produces the tumor protein p53) in the progression to cancer. The protein p53 has been described as "the guardian of the genome" because of its role in conserving stability by preventing genome mutations. TP53 has since been implicated in many cancers, see [11] and [25].

Combining the ideas in the last two paragraphs leads to a four (or five) stage description for colon cancer that is described for example in the books of Vogelstein and Kinzel [24], and Frank [9]. In 2002, Leubeck and Moolgavar [16] developed a mathematical model in order to fit the age-incidence of colorectal cancer. We will describe the model in detail in the next section. They tried models with $k = 2, 3, 4, 5$ stages and found that the four-stage model gave the best fit. The techniques developed in [16] have been applied to study a number of other cancers. See e.g. [17], [18], [19].

2 Analytic approach

In the k-stage model there is a fixed number of stem cells, N, each of which mutates at rate μ_0 to become a type 1 cell, so cells of type 1 are born at times of a Poisson process with rate $\gamma = N\mu_0$. Cells of types $i = 1, \ldots k-2$ are pre-initiated cells that mutate at rate

$\mu_{i,k}$ to become a cell of type $i+1$. Cells of type $k-1$ are initiated cells that divide into two at rate α, die at rate β, where $\lambda = \alpha - \beta > 0$, and mutate at rate $\mu_{k-1,k}$ to become malignant (type k). Let $Z_i^k(t)$ be the number of type i cells at time t in the k-stage model. Let $T_k = \min\{t : Z_k^k(t) > 0\}$ be the time of appearance of the first malignant cell in the k-stage model. Here we will be interested primarily in

- the survival function $H_k(t) = P(T_k > t)$, which gives the fraction of individuals that are cancer free at time t
- hazard rate $h_k(t) = -H_k'(t)/H_k(t) = \frac{d}{dt} \ln H_k(t)$, which gives the rate at which healthy individuals become sick at time t.

The traditional approach to studying the k-stage model, as explained for example in the supplementary materials of [16], has been to let

$$P(i_1, i_2, \dots i_k; t) = P\left(Z_1(t) = i_1, Z_2(t) = i_2, \dots Z_k(t) = i_k\right),$$

define the generating function

$$\Psi_k(y_1, y_2, \dots y_k; t) = \sum_{i_1, i_2, \dots i_k} P(i_1, i_2, \dots i_k; t) y_1^{i_1} y_2^{i_2} \cdots y_k^{i_k},$$

and compute Ψ_k by solving a hyperbolic PDE using the method of characteristics.

To explain this approach, we will consider the case $k = 2$ and write

$$\Psi(y, z; t) = \sum_{j,k} P(Z_1(t) = j, Z_2(t) = k) y^j z^k.$$

Transition rates are given in the following table.

at rate	transition	g.f. change
γ	$j \to j+1$	$\Psi \to y_1 \Psi$
αj	$j \to j+1$	$\Psi \to y_1 \Psi$
βj	$j \to j-1$	$\Psi \to \Psi / y_1$
$\mu_{1,2} j$	$k \to k+1$	$\Psi \to y_2 \Psi$

From the table we get

$$\frac{\partial \Psi}{\partial t} = \gamma(y_1 - 1)\Psi + \alpha j(y_1 - 1)\Psi + \beta j(1/y_1 - 1)\Psi + \mu_{1,2} j(y_2 - 1)\Psi.$$

Using the identity $j\Psi = y_1 \partial\Psi/\partial y_1$ this becomes

$$\frac{\partial \Psi}{\partial t} = \gamma(y_1 - 1)\Psi + \left[\alpha(y_1 - 1) + \beta(1/y_1 - 1) + \mu_{1,2}(y_2 - 1)\right] y_1 \frac{\partial \Psi}{\partial y_1},$$

which rearranges to become

$$\frac{\partial \Psi}{\partial t} = \gamma(y_1 - 1)\Psi + \left[\alpha y_1^2 - (\alpha + \beta + \mu_{1,2}(1 - y_2))y_1 + \beta\right] \frac{\partial \Psi}{\partial y_1}. \tag{2}$$

To find the generating function $\Psi_k(z_1, z_2, \ldots z_k : t)$, one uses the fact that the solution is constant along characteristic curves, i.e.,

$$\Psi_k(z_1, z_2, \ldots z_k; t) = \Psi(y_{1,k}(s,t), y_{2,k}(s,t), \ldots y_{k,k}(s,t); s)$$

where the $y_{i,k}(s,t)$ satisfy the characteristic equations

$$\begin{aligned} y'_{k,k}(s,t) &= 0 \\ y'_{k-1,k}(s,t) &= \alpha y_{k-1,k}^2 - (\alpha + \beta + \mu_{k-1,k}(1 - y_{k,k}))y_{k-1,k} + \beta \\ y'_{i,k}(s,t) &= -\mu_{i,k}(1 - y_{i+1,k}(s,t))y_{i,k}(s,t) \qquad 1 \le i \le k-2 \end{aligned} \tag{3}$$

and the derivative is taken with respect to the s variable. The generating function can then be found from

$$\Psi_k(z_1, z_2 \ldots z_k : t) = \exp\left(-\gamma \int_0^t 1 - y_{1,k}(s,t)\, ds\right) \tag{4}$$

and the survival function $H_k(t) = P(Z_k^k(t) = 0) = \Psi_k(1, \ldots, 1, 0; t)$.

2.1 Solving the equations

To begin to solve the equations in (3), we note that $y'_{k,k}(s,t) = 0$ so $y_{k,k}(s,t) = 0$ is constant. We write $S_{1,k}(t) = y_{k-1,k}(t)$ since it is the first step in solving the system of equations. The subscript k is needed because the mutation rates that enter into the differential equations (3) depend on the number of stages. Throughout this paper we when we write $S_{i,k}(t)$ it is assumed that $z_k = 0$ and $z_i = 1$ for $1 \le i < k$. Changing notation we want to solve

$$S'_{1,k}(t) = \alpha S_{1,k}^2 - (\alpha + \beta + \mu)S_{1,k}(t) + \beta, \tag{5}$$

where $\mu = \mu_{k-1,k}$ and $S_{1,k}(0) = 1$. The quadratic equation $\alpha x^2 - (\alpha + \beta + \mu)x + \beta = 0$ has two roots $q > 1 > r > 0$. See (31). Solving (5), see (34), gives

$$S_{1,k}(t) = r + \frac{q - r}{1 + \frac{q-1}{1-r}\exp(\alpha(q-r)t)} \tag{6}$$

Having solved for $S_{1,k}(t)$ the other $S_{i,k}(t) = y_{k-i,k}(t)$ can be found by induction:

$$S_{i,k}(t) = \exp\left(-\nu_i \int_0^t (1 - S_{i-1,k}(t-s))\, ds\right) \tag{7}$$

where in the k-stage model $\nu_i = \mu_{k-i,k}$. The computation of the $S_{i,k}(t)$ is not easily found in the literature, so we will give the details in Section 7.

While the recursion in (7) was derived by the method of characteristics, it has a simple probabilistic interpretation. Each individual of type $k-i$ gives birth to individuals of type $k-i+1$ at times of a rate ν_i Poisson process. A type $k-i+1$ born at time s will give rise to a malignant cell with probability $1 - S_{i-1}(t-s)$. The number of type $k-i+1$ individuals that are successful in doing this has a Poisson distribution with mean

$$\rho_{i,k}(t) = \nu_i \int_0^t 1 - S_{i-1,k}(t-s)\, ds \tag{8}$$

so the probability none of the type $k-i+1$ individuals are successful in creating a malignant cell is $S_{i,k}(t) = \exp(-\rho_{i,k}(t))$.

2.2 Hazard rate formulas

It is clear from (4) that $H_k = S_{k,k}$ so we have

$$H_k(t) = \exp\left(-\gamma \int_0^t 1 - S_{k-1,k}(t-s)\, ds\right) \tag{9}$$

Using (9) and changing variables $r = t - s$ before differentiating it follows that

$$h_k(t) = -\frac{H_k'(t)}{H_k(t)} = \gamma(1 - S_{k-1,k}(t)) \tag{10}$$

so we do not have to evaluate $H_k(t)$ to find $h_k(t)$.

Using (10) with (35) gives

$$h_2(t) = \gamma \cdot \frac{(q-1)(1-r)e^{-\alpha(r-1)t} - (q-1)(1-r)e^{-\alpha(q-1)t}}{(q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}} \tag{11}$$

Using (37) we have

$$h_3(t) = \gamma\left(1 - \left[\frac{q-r}{(q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}}\right]^{\mu_1/\alpha}\right) \tag{12}$$

When it comes to the fourth stage, the possibility to compute the integral in (7) breaks down and (39) gives

$$h_4(t) = \gamma\left(1 - \exp\left(-\mu_1 \int_0^t 1 - \left[\frac{q-r}{(q-1)e^{-\alpha(r-1)u} - (r-1)e^{-\alpha(q-1)u}}\right]^{\mu_2/\alpha} du\right)\right) \tag{13}$$

3 Probabilistic approach

We begin by giving probabilist interpretations for some of the computations above. To explain the differential equation for $y_{k-1,k}$, note that if we ignore mutation then each individual of type $k-1$ initiates a linear birth and death process $L(t)$ in which the number of individuals increases from $m \to m+1$ at rate αm and decreases from $m \to m-1$ at rate βm, where $\alpha > \beta$.

Theorem 1. *As $t \to \infty$, $e^{-\lambda t}L(t) \to W$ with $P(W = 0) = \beta/\alpha = P(L(t) = 0$ for some $t > 0)$ and*

$$P(W > x | W > 0) = \exp(-x\lambda/\alpha)$$

i.e., if we condition on non-extinction then W has an exponential density with rate λ/α. If we let $V_0 = (W|W > 0)$ then

$$E(e^{-\theta V_0}) = \int_0^\infty \frac{\lambda}{\alpha} e^{-x\lambda/\alpha} e^{-\theta x}\, dx = \frac{\lambda/\alpha}{\lambda/\alpha + \theta} \tag{14}$$

Proof. We will sketch the proof since it contains details that will be useful later. For more details see Section 3 of [6]. It is well know that if we start with $L(0) = 1$ then the generating function $F(x,t) = Ex^{L(t)}$ satisfies

$$\frac{\partial F}{\partial t} = -(\alpha+\beta)F + \alpha F^2 - \beta \tag{15}$$

with boundary condition $F(x,0) = x$. This equation can be solved with the result that

$$F(x,t) = \frac{\beta(x-1) - e^{\lambda t}(\alpha x - \beta)}{\alpha(x-1) - e^{\lambda t}(\alpha x - \beta)}$$

where $\lambda = \alpha - \beta$ is the exponential growth rate. By considering what happens on the first step (which is a birth with probability $\alpha/(\alpha+\beta)$ and a death with probability $\beta/(\alpha+\beta)$) we can conclude that the probability ρ that the process dies out satisfies

$$\rho = \rho^2 \cdot \frac{\alpha}{\alpha+\beta} + 1 \cdot \frac{\beta}{\alpha+\beta}$$

The extinction probability is the root which is < 1, i.e., $\rho = \beta/\alpha$. □

Comparing with (15) we see that (5) has an additional term $-\mu y(t)$. In probabilistic terms, this corresponds to killing the process at rate μm when there are m individuals in the branching process. Let $\bar{L}_t$ be the birth and death chain conditioned not to die out. Using this observation and Theorem 1, the probability of no malignant cell by time t in $\bar{L}_t$ is

$$\begin{aligned} G(t) &= E\exp\left(-\int_0^t \mu \bar{L}(s)\,ds\right) \approx E\exp\left(-\mu\int_0^t e^{\lambda s}V_0\,ds\right) \\ &\approx E\exp(-\mu e^{\lambda t}V_0/\lambda) = \frac{\lambda/\alpha}{\lambda/\alpha + \mu e^{\lambda t}/\lambda} \end{aligned} \tag{16}$$

where in the last step we have used (14). When the branching process dies out it does that quickly so the probability of a mutation is small. From this it follows that

$$S_{1,k}(t) = \frac{\beta}{\alpha} + \frac{\lambda}{\alpha}\cdot\frac{\lambda/\alpha}{\lambda/\alpha + \mu e^{\lambda t}/\lambda} = \frac{\beta}{\alpha} + \frac{\lambda/\alpha}{1 + \mu\alpha e^{\lambda t}/\lambda^2} \tag{17}$$

To compare with (6) we need to change notation. Equations (41) and (42) imply that

$$q \approx 1 + \frac{\mu}{\lambda} \qquad r \approx \frac{\beta}{\alpha} \qquad q - r \approx \frac{\lambda}{\alpha} \tag{18}$$

Using these in (6) we have

$$S_{1,k}(t) \approx \frac{\beta}{\alpha} + \frac{\lambda/\alpha}{1 + (\mu/\lambda)(\alpha/\lambda)e^{\lambda t}}$$

which agrees with the new formula in 17.

To compute the survival function $H_k(t)$ from this we start at 1 and work up to type $k-1$. Let $\eta_j(s)$ be the rate at which type j's born at time s. Integrating we find

$$\begin{aligned}
\eta_1(s) &= \gamma = N\mu_0 \\
\eta_2(t) &= \mu_1 \int_0^t \eta_1(s)\, ds = \gamma\mu_1 s \\
\eta_3(t) &= \mu_2 \int_0^t \eta_2(s)\, ds = \gamma\mu_1\mu_2 s^2/2! \\
\eta_j(t) &= \mu_{j-1} \int_0^t \eta_{j-1}(s)\, ds = \gamma\mu_1 \ldots \mu_{j-1} s^{j-1}/(j-1)!
\end{aligned}$$

We call a type $k-1$ family that does not die out "successful." Using Theorem 1, the probability that a type $k-1$ is successful is λ/α. On the event that a type $k-1$ is produced in $[0,T]$, the time it is born will be distributed as $\eta_{k-1}(s)\lambda/\alpha$. Recalling that $1-G(t-s)$ is the probability a successful birth and death process conditioned to not die out gives rise to a malignant cell and using the reasoning that led to (8), the times of successes will be roughly a Poisson process so

$$H_k(t) \approx \exp\left(-\int_0^t \eta_{k-1}(t)\cdot\frac{\lambda}{\alpha}\cdot(1-G(t-s))\, ds\right). \tag{19}$$

Using this approach we find (see Section 8 for details) that

$$\begin{aligned}
H_2(t) &= \exp\left(-\frac{\gamma}{\alpha}\left[\ln\left(\frac{\mu_1}{\lambda}e^{\lambda t}+\frac{\lambda}{\alpha}\right)-\ln(\lambda/\alpha)\right]\right) \\
H_3(t) &= \exp\left(-\gamma\frac{\mu_1}{\alpha}\int_0^t \log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda(u)})\, du\right) \\
H_4(t) &= \exp\left(-\gamma\mu_1\frac{\mu_2}{\alpha}\int_0^t (t-u)\log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u})\, du\right)
\end{aligned}$$

Differentiating with respect to t, and using $h_k(t) = -H_k'(t)/H_k(t)$

$$h_2(t) = \frac{\gamma}{\alpha}\cdot\frac{\mu_1 e^{\lambda t}}{(\mu_1/\lambda)e^{\lambda t}+(\lambda/\alpha)} \tag{20}$$

$$h_3(t) = \gamma\frac{\mu_1}{\alpha}\log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u}) \tag{21}$$

$$h_4(t) = \gamma\mu_1\int_0^t \frac{\mu_2}{\alpha}\log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u})\, du \tag{22}$$

At first glance the new formulas look much different than the ones from the analytic approach. However, a closer look shows they are closely related. When $k=2$, $\eta_{k-1}=\gamma$ and (17) implies

$$\frac{\lambda}{\alpha}(1-G(t)) = 1 - S_{1,2}(t)$$

so the formulas for $h_2(t)$ agree. Comparing the two formulas for $h_3(t)$ and $h_4(t)$ we see that it is enough to argue

$$\frac{\mu_i}{\alpha}\log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u}) \approx 1-\left[\frac{q-r}{(q-1)e^{-\alpha(r-1)u}-(r-1)e^{-\alpha(q-1)u}}\right]^{\mu_i/\alpha} \tag{23}$$

4 The two-stage model

.

The formula for the hazard rate in the two stage case, given in (11), is

$$h_2(t) = \gamma(q-1)(1-r)\frac{e^{-\alpha(r-1)t} - e^{-\alpha(q-1)t}}{(q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}}$$

If we introduce $P = \alpha(r-1)$ and $Q = \alpha(q-1)$ we can rewrite it as

$$\begin{aligned} &= \frac{\gamma}{\alpha}\alpha(q-1)\alpha(1-r)\frac{e^{-\alpha(r-1)t} - e^{-\alpha(q-1)t}}{\alpha(q-1)e^{-\alpha(r-1)t} - \alpha(r-1)e^{-\alpha(q-1)t}} \\ &= \frac{\gamma}{\alpha}\frac{(-PQ)(e^{-Pt} - e^{-Qt})}{Qe^{-Pt} - Pe^{-Qt}} \end{aligned} \tag{24}$$

Note that the two-stage model has five parameters γ, μ_1, μ_2, α, and β but the new formula for the hazard rate has only three γ/α, $P = \alpha(r-1)$, and $Q = \alpha(q-1)$. In the terminology of statistics we have an identifiability problem, i.e., not all the parameters in the model can be estimated.

parameter	[18]	[17]
γ/α	3.87×10^{-4}	3.17×10^{-5}
$-P$	0.259	0.11
Q	8.22×10^{-4}	$1.78 \times 10^{-}$

Table 1: Fits of the two stage model hazard rate given in (24) to thyroid cancer in white females by Meza and Chang [18] and to peritoneal mesothelioma by Moolgavkar, Meza, and Turim [17].

Figure 1 gives a picture of $h_2(t)$ for the parameters of [18]. It should be obvious from the picture that as $t \to \infty$, $h_2(t)$ converges to a limit. Since $P < 0 < Q$ letting $t \to \infty$ in (24)

$$h_2(t) \to \frac{\gamma}{\alpha}(-P) = \gamma(1-r). \tag{25}$$

Figure 2 shows the fit of the two stage model to peritoneal mesotheliomas in SEER data from 1973–2005. Parameters are given in Table 1. This time the asymptote has not been reached by age 85. The dotted line gives the fit of the Armitage-Doll formula (1) Ct^k to the data with $C = 1.75 \times 10^{-11}$ and $k = 2.79$. Visually the second fit is worse. This is confirmed by the values of Akaike Information Criterion scores. Interested readers can consult [17] for further details.

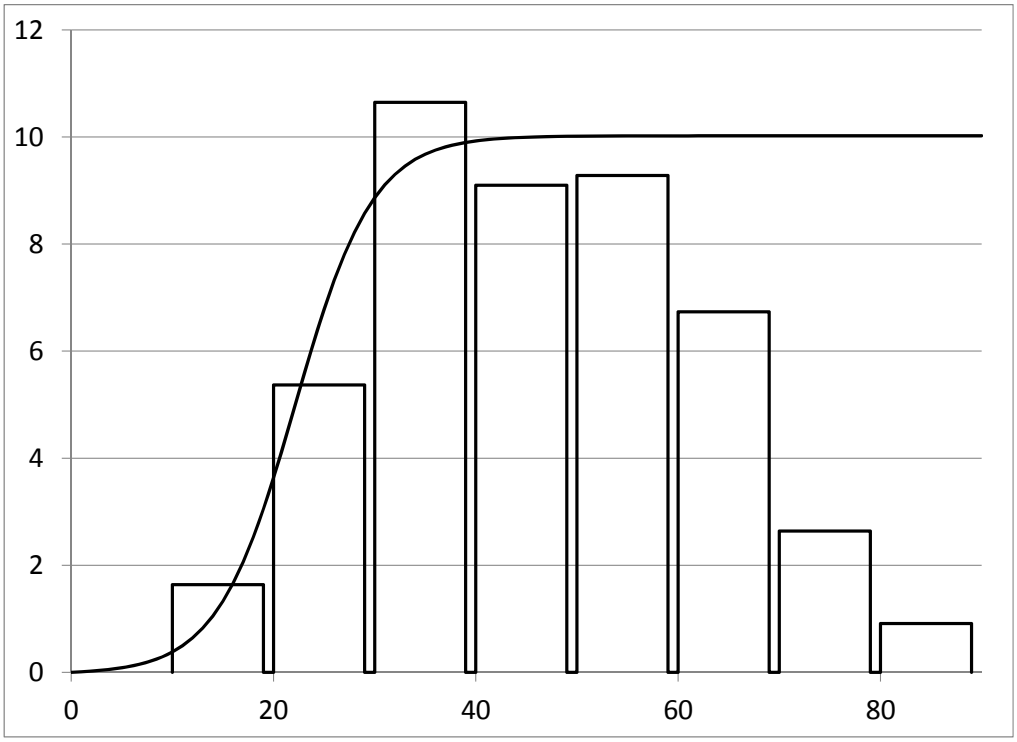

Figure 1: A graph of $h_2(t)$ for the thyroid cancer parameters. x axis is age in years. y axis is cases per 100,000 per year. The asymptotic value, which is 10.02 by (25) is reached at age ≈ 40. For comparison, we give a histogram of age at diagnosis in 508 individuals in a TCGA study [22]. If one transforms the data so that it is cases per 100,000 individuals in each age group, the model fits the data. See Figure 4 in [18].

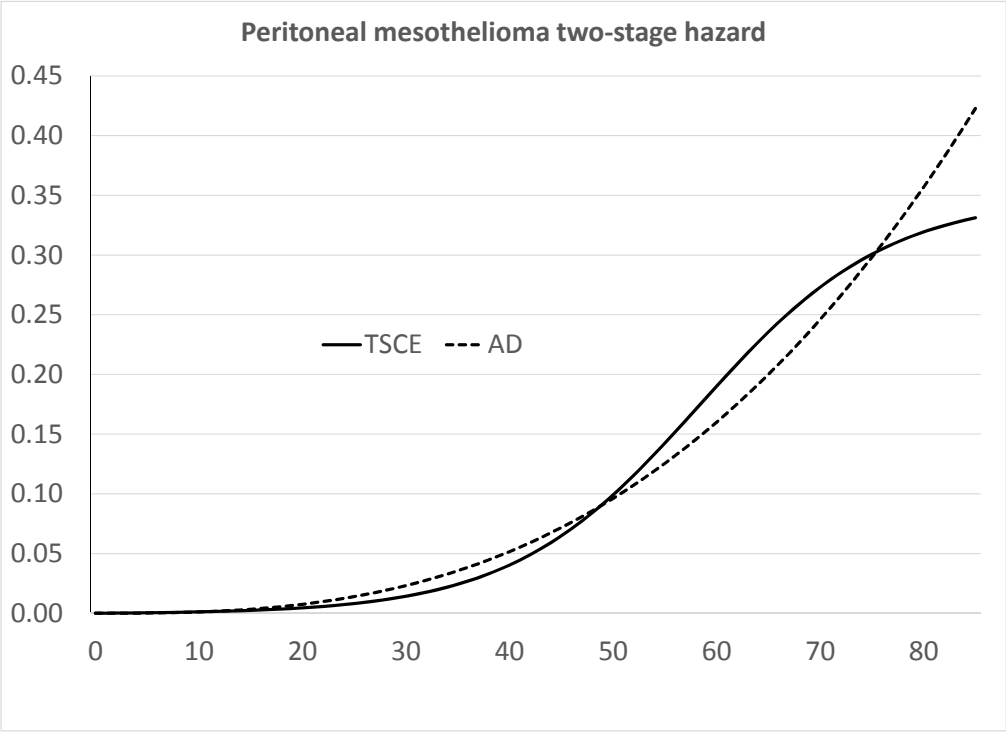

Figure 2: The solid line is a graph of $h_2(t)$ for the peritoneal mesothelioma parameters. The dotted line is a fit of the Armitage-Doll model with $k = 2.79$. x axis is age in years. y axis is cases per 100,000 per year. The asymptotic value has not been reached by age 85.

5 The three-stage model

Using (12) and changing variables $P = \alpha(r-1)$ and $Q = \alpha(q-1)$ it follows that the hazard rate is

$$h_3(t) = \gamma\left(1 - \left[\frac{Q-P}{Qe^{-Pt} - Pe^{-Qt}}\right]^{\mu_1/\alpha}\right) \tag{26}$$

Meza at al [19] show that h_3 is asymptotically linear. To state their results we need two definitions. The probability that the birth and death processes does not die out is

$$p_\infty = \lim_{t\to\infty} 1 - S_1(t) = 1 - r \approx 1 - \beta/\alpha.$$

by (6) and (18). Let $T_{2,3}$ be the time to malignancy of a single type 2 clone in the 3-stage model conditional on it not becoming extinct.

Theorem 2. *If t is large and $t \ll 1/\mu_1 p_\infty$ then*

$$h_3(t) \approx \gamma\mu_1 p_\infty(t - ET_{2,3}) \quad \textit{where} \quad ET_{2,3} \approx -\frac{1}{\alpha-\beta}\ln\left(\frac{\mu_2\alpha}{(\alpha-\beta)^2}\right) \tag{27}$$

To better understand the formula for $h_3(t)$, and to check the accuracy of the linear approximation, it is useful to have concrete examples.

name	parameter	3-stage fit	4-stage fit
slope	$\gamma\mu_1 p_\infty$	3.9×10^{-5}	4.68×10^{-5}
$-P$	$\alpha - \beta$	0.179	0.192
	μ_2/α	NA	0.401
Q	$\alpha\mu_{k-1}/(\alpha-\beta)$	1.38×10^{-5}	2.06×10^{-5}
		$ET_{2,3} = 52.9$	$ET_{2,4} = 57.9$

Table 2: Meza et al [19] estimated parameters for the 3-stage model and 4-stage model (described in the next section) for pancreatic cancer in men.

To be able to compute the hazard function we need a value for α. Meza at al [19] suggest $\alpha = 9$ cell divisions per year and say that the fit is not sensitive to the value of α chosen. In pancreatic cancer, when $\alpha = 9$,

$$\mu_2 = -QP/\alpha = 2.74 \times 10^{-7}, \quad p_\infty = (\alpha-\beta)/\alpha \approx 0.02 \quad N\mu_0\mu_1 \approx 2 \times 10^{-3} \tag{28}$$

If we take $\mu_0 = \mu_1 = 10^{-6}$ then $N = 2 \times 10^9$, and $\mu_1/\alpha = 1.1 \times 10^{-7}$. Figure 3 gives a graph of $h_3(t)$ (and $h_4(t)$) for the pancreatic cancer parameters. $1/\mu_1 p_\infty = 5 \times 10^9$ years so the condition $t \ll 1/\mu_1 p_\infty$ holds. As the graph shows the straight line approximation is good for $t \geq 65$.

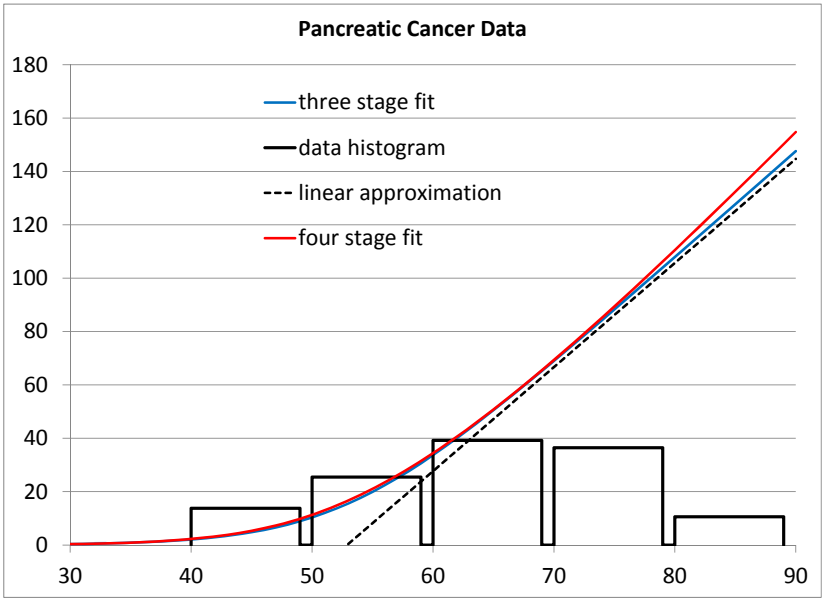

Figure 3: Graphs of $h_3(t)$ and $h_4(t)$ for the pancreatic cancer parameters. x axis is age in years. y axis is cases per 100,000 per year. Straight line is the linear approximation to the three stage model (27). The bar graph gives the age at diagnosis for 186 patients in the TCGA study of pancreatic cancer [4]. Again if one transforms the data to be cases per 100,000 in each age group, the theoretical curve fits the data, see Figure 5 in [19].

6 The four-stage model

6.1 Hazard rate

Using (13) and changing variables $P = \alpha(r-1)$ and $Q = \alpha(q-1)$

$$h_4(t) = \gamma\left(1 - \exp\left(-\mu_1 \int_0^t 1 - \left[\frac{Q-P}{Qe^{-P(t-s)} - Pe^{-Q(t-s)}}\right]^{\mu_2/\alpha} ds\right)\right) \tag{29}$$

Let $T_{2,4}$ be the time for a single type 2 clone to produce a malignant cell in the four-stage model and let

$$p_\infty = \lim_{t\to\infty} 1 - S_{2,4}(t) = 1$$

since a type 2 will give rise to infinitely many type 3's, and one of these will start a branching process that does not die out.

The asymptotic behavior of the hazard rate is constant in the two-stage case and linear in the three-stage case. One might naively guess that in the four-stage case it is asymptotically quadratic, but the simple proof given below shows it is asymptotically linear. It should be clear from the proof that this holds for any $k \geq 3$.

Theorem 3. *When t is large and $t \ll 1/\mu_1$*

$$h_4(t) \approx \gamma\mu_1(t - ET_{2,4})$$

Proof of Theorem 3. When $\mu_1 t$ is small

$$\begin{aligned} h_4(t) &= \gamma(1 - S_{3,4}(t)) = \gamma(1 - e^{-\mu_1 \int_0^t (1 - S_{2,4}(t-s)\, ds}) \\ &\approx \gamma\mu_1 \int_0^t 1 - S_{2,4}(t-s)\, ds = \mu_1 \left(t - \int_0^t S_{2,4}(t)\, dt\right) \end{aligned}$$

Using a well-known formula for expected value

$$ET_{2,4} = \int_0^\infty P(T_{2,4} \geq t)\, dt = \int_0^\infty S_{2,4}(t)\, dt$$

Combining the last two equations gives the desired result. □

Note: due to the complexity of the formula for $S_{2,4}(t)$ given in (36), we do not have a formula for $ET_{2,4}$. However, it is easy to compute numerically.

Leubeck and Moolgavkar [16] estimated parameters for the 2, 3, 4, and 5 stage models for colorectal cancer in women. As Figure 4 shows the fits from the four models are all very good. To explain how this could happen, we take a look at the parameters used in fitting. $N = 10^8$, $\alpha = 9$. Note that in the four stage model, $\mu_2 = 6.3$, and in the five-stage model $\mu_3 = \mu_4 = 0.9$. These large values speed these processes up, effectively eliminating one and two stages respectively. In the other direction in the two stage model the very slow mutation rates $\mu_0 = 4.5 \times 10^{-9}$ and $\mu_1 = 1.44 \times 10^{-7}$ effectively add a stage. Thus if we judge the fitted model by the size of the mutation rates, it seems that the three-stage model gives the best fit.

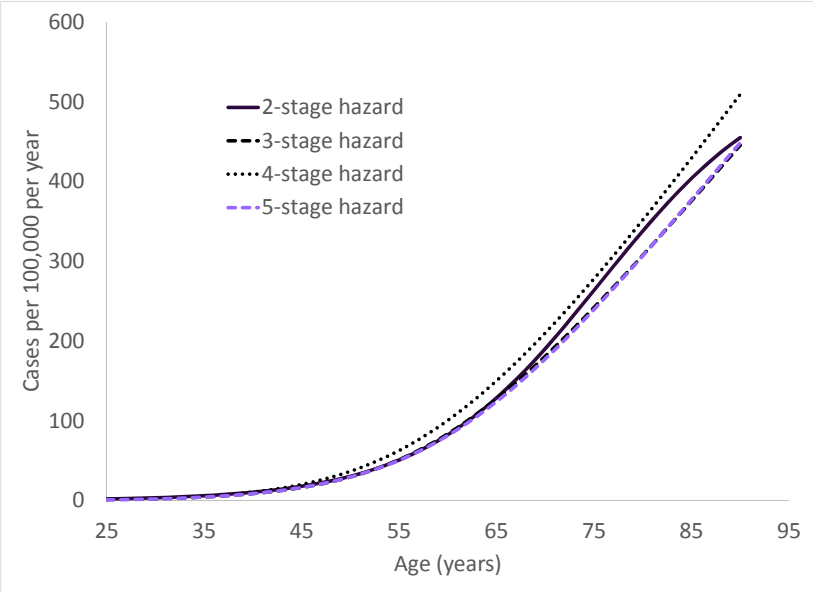

Figure 4: A comparison of the fitted values of the hazard functions for the two, three, four, and five stage models of [16]. The three and five stage fits are almost identical so you can only see three curves on the graph.

	2-stage	3-stage	4-stage	5-stage
μ_0	4.5×10^{-9}	1×10^{-5}	1.3×10^{-6}	1.3×10^{-6}
μ_1	1.44×10^{-7}	1×10^{-5}	1×10^{-6}	1×10^{-6}
μ_2	—	8.77×10^{-7}	6.3	0.9
μ_3	—	—	1.333×10^{-6}	0.9
μ_4	—	—	—	1.89×10^{-6}
P	-0.11	-0.13	-0.13	-0.11
Q	2.64×10^{-5}	6.08×10^{-5}	9.23×10^{-3}	1.545×10^{-4}

Table 3: Parameter values in four fits of colon cancer data from [16].

It is interesting to note that Tomasetti et al [23] have arrived at the conclusion colon cancer is a three-stage process by a completely different reasoning. They compared patients with and without a mismatch repair deficiency. They found that the latter group has 7.7 to 8.8 times as many mutations, versus a 114.2 fold increase in colon cancer rates, and argued that the increase would be more substantial if the process had four-stages. See pages 119–120 in [23] for more details and an analysis of lung adenocarcinomas.

7 Computing $S_{i,k}(t)$: analytic approach

Let $q > 1 > r$ be the roots of the quadratic equation

$$\alpha y^2 - (\alpha + \beta + \mu)y + \beta = 0 \tag{30}$$

that is,

$$\frac{\alpha + \beta + \mu \pm \sqrt{(\alpha + \beta + \mu)^2 - 4\alpha\beta}}{2\alpha}. \tag{31}$$

If we write $y(t) = S_{1,k}(t)$ then the differential equation (5) can be written as

$$y'(t) = \alpha(y - q)(y - r) \qquad y(0) = 1 \tag{32}$$

From this we see that $y(t)$ is decreasing and will converge to r as $t \to \infty$. Rearranging (32), we have

$$\alpha\, ds = \frac{dy}{(y - q)(y - r)} = -\frac{1}{q - r}\left(\frac{dy}{q - y} - \frac{dy}{y - r}\right).$$

Here we have written the right-hand side to avoid taking the logarithm of a negative number in the next step. Multiplying both sides by $q - r$ and then integrating from 0 to t, we have for some constant D

$$\alpha(q - r)t + D = \log(q - y) - \log(y - r) = \log\left(\frac{q - y}{y - r}\right).$$

Exponentiating we have

$$q - y = (y - r)\exp(\alpha(q - r)t + D). \tag{33}$$

Solving for y gives

$$y(t) = \frac{q + r\exp(\alpha(q - r)t + D)}{1 + \exp(\alpha(q - r)t + D)}$$

Using (33) and recalling $y(t) = 1$ we have $e^D = (q - 1)/(1 - r)$ which implies

$$\begin{aligned} y(t) &= \frac{q + r\frac{1-q}{1-r}\exp(\alpha(q - r)t)}{1 + \frac{1-q}{1-r}\exp(\alpha(q - r)t)} \\ &= r + \frac{q - r}{1 + \frac{q-1}{1-r}\exp(\alpha(q - r)t)}. \end{aligned} \tag{34}$$

which is (6). Our next step is to write

$$\begin{aligned} 1 - S_{1,k}(t) &= 1 - r - \frac{q - r}{1 + \frac{q-1}{1-r}\exp(\alpha(q - r)t)} \\ &= \frac{1 - q + (q - 1)\exp(\alpha(q - r)t)}{1 + \frac{q-1}{1-r}\exp(\alpha(q - r)t)} \\ &= \frac{(1 - q)(1 - r)e^{\alpha(1-q)t} + (1 - r)(q - 1)e^{\alpha(1-r)t}}{(1 - r)e^{\alpha(1-q)t} + (q - 1)e^{\alpha(1-r)t}} \\ &= \frac{(q - 1)(1 - r)e^{-\alpha(r-1)t} - (q - 1)(1 - r)e^{-\alpha(q-1)t}}{(q - 1)e^{-\alpha(r-1)t} - (r - 1)e^{-\alpha(q-1)t}} \end{aligned} \tag{35}$$

To compute $S_{2,k}(t)$ we use the recursion (7)

$$S_{2,k}(t) = \exp\left(-\nu_2 \int_0^t (1 - S_{1,k}(s))\, ds\right).$$

To compute integral let $f(t) = (q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}$ and note that

$$f'(t) = \alpha(q-1)(1-r)e^{-\alpha(r-1)t} - \alpha(q-1)(1-r)e^{-\alpha(q-1)t}$$

so we have

$$\begin{aligned} S_{2,k} &= \exp\left(-\nu_2 \int_0^t \frac{f'(t)}{\alpha f(t)}\right) = \exp\left(-\frac{\nu_2}{\alpha}\log(f(t)/f(0))\right) \\ &= \left[\frac{q-r}{(q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}}\right]^{\nu_2/\alpha}. \end{aligned} \tag{36}$$

When $k = 3$, $\nu_2 = \mu_1$ so we have

$$h_3(t) = \gamma(1 - S_{2,3}(t)) = \gamma\left(1 - \left[\frac{q-r}{(q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}}\right]^{\mu_1/\alpha}\right). \tag{37}$$

Integrating again we conclude that

$$\begin{aligned} S_{3,k}(t) &= \exp\left(-\nu_3 \int_0^t 1 - S_{2,k}(t-s)\, ds\right) \\ &= \exp\left(-\nu_3 \int_0^t 1 - \left[\frac{q-r}{(q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}}\right]^{\nu_2/\alpha} ds\right). \end{aligned} \tag{38}$$

When $k = 4$, $\nu_3 = \mu_1$ and $\nu_2 = \mu_2$ so

$$h_4(t) = \gamma\left(1 - \exp\left(-\nu_3 \int_0^t 1 - \left[\frac{q-r}{(q-1)e^{-\alpha(r-1)t} - (r-1)e^{-\alpha(q-1)t}}\right]^{\nu_2/\alpha} ds\right)\right). \tag{39}$$

7.1 Approximations for q and r

When $\mu = 0$ the roots q, r are (31)

$$q = \frac{\alpha + \beta \pm \sqrt{(\alpha-\beta)^2}}{2\alpha} = 1 \quad \text{and} \quad r = \frac{\beta}{\alpha} \tag{40}$$

Typically the mutation rate μ is much smaller than α and β. When it is

$$(\alpha+\beta+\mu)^2 - 4\alpha\beta \approx (\alpha+\beta)^2 + 2(\alpha+\beta)\mu - 4\alpha\beta = (\alpha-\beta)^2 + 2(\alpha+\beta)\mu$$

so we have

$$\sqrt{(\alpha+\beta+\mu)^2 - 4\alpha\beta} \approx (\alpha-\beta) + \frac{(\alpha+\beta)}{(\alpha-\beta)}\mu,$$

and it follows that

$$q = 1 + \frac{\mu}{2\alpha} + \frac{\alpha+\beta}{2\alpha(\alpha-\beta)} = 1 + \frac{\mu}{\alpha-\beta}, \tag{41}$$

$$r = \frac{\beta}{\alpha} + \frac{\mu}{2\alpha} - \frac{\alpha+\beta}{2\alpha(\alpha-\beta)} = \frac{\beta}{\alpha}\left(1 - \frac{\mu}{\alpha-\beta}\right) \approx \frac{\beta}{\alpha}. \tag{42}$$

8 Hazard functions $H_k(t)$ probabilistic approach

Using (19) with the formula for $G(t)$ given in (16)

$$\begin{aligned} H_2(t) &= \exp\left(-\frac{\gamma}{\alpha}\int_0^t \frac{\mu_1 e^{\lambda(t-s)}}{\lambda/\alpha + \mu_1 e^{\lambda(t-s)}/\lambda}\,ds\right) \\ &= \exp\left(-\frac{\gamma}{\alpha}\left[\ln\left(\frac{\mu_1}{\lambda}e^{\lambda t} + \frac{\lambda}{\alpha}\right) - \ln(\lambda/\alpha)\right]\right) \end{aligned} \tag{43}$$

If we differentiate (43) we get

$$h_2(t) = -\frac{H_2'(t)}{H_2(t)} = \frac{\gamma}{\alpha}\cdot\frac{\mu_1 e^{\lambda t}}{(\mu_1/\lambda)e^{\lambda t} + (\lambda/\alpha)}.$$

Theorem 4.

$$H_3(t) = \exp\left(-\gamma\frac{\mu_1}{\alpha}\int_0^t \log(1 + (\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u})\,du\right)$$

and differentiating gives

$$h_3(t) = -\frac{H_3'(t)}{H_3(t)} = -\gamma\frac{\mu_1}{\alpha}\log(1 + (\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u}).$$

Proof. Using (19) with the formula for $G(t)$ given in (16)

$$\begin{aligned} H_3(t) &= \exp\left(-\gamma\frac{\mu_1}{\alpha}\int_0^t s\cdot\frac{\mu_2 e^{\lambda(t-s)}}{\lambda/\alpha + \mu_2 e^{\lambda(t-s)}/\lambda}\,ds\right) \\ &= \exp\left(-\gamma\frac{\mu_1}{\alpha}\int_0^t s\cdot\frac{(\alpha/\lambda)\mu_2 e^{\lambda(t-s)}}{1 + (\alpha/\lambda)(\mu_2/\lambda)e^{\lambda(t-s)}}\,ds\right). \end{aligned} \tag{44}$$

Integrating by parts with $f(s) = s$ and $g'(s) =$ the fraction under the integral

$$H_3(t) = \exp\left(-\gamma\frac{\mu_1}{\alpha}\int_0^t \log[1 + (\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda(t-s)}]\,ds\right) \tag{45}$$

since $f(s)g(s) = 0$ when $s = 0$ and $s = t$. now change variables $u = t - s$. □

Proof of (23). If x is small then $x \approx 1 - e^{-x}$. Using this with

$$x = \frac{\mu_i}{\alpha}\log(1 + (\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda(u)})$$

we have

$$1 - e^{-x} = 1 - \left[\frac{1}{1 + (\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u}}\right]^{\mu_i/\alpha} = 1 - \left[\frac{\lambda/\alpha}{\lambda/\alpha + (\mu_2/\lambda)\theta e^{\lambda u}}\right]^{\mu_i/\alpha}$$

Using (18) $q - r \approx \lambda$, and $q - 1 \approx \mu_2/\lambda$, so the above is

$$\approx \left[\frac{q-r}{(q-r) + (q-1)e^{\alpha(q-r)t}}\right]^{\mu_i/\alpha} \approx \left[\frac{q-r}{(1-r)e^{-\alpha(q-1)t} + (q-1)e^{-\alpha(r-1)t}}\right]^{\mu_i/\alpha}$$

proving the desired result. □

Theorem 5.

$$H_4(t) = \exp\left\{-\gamma\mu_1\frac{\mu_2}{\alpha}\int_0^t (t-u)\log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u})\,du\right\}$$

Differentiating with respect to t

$$h_4(t) = -\frac{H_4'(t)}{H_4(t)} = \gamma\mu_1\frac{\mu_2}{\alpha}\int_0^t \log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda u})\,ds \tag{46}$$

Proof. Using (19) with the formula for $G(t)$ given in (16)

$$\begin{aligned} H_4(t) &= \exp\left(-\gamma\mu_1\frac{\mu_2}{\alpha}\int_0^t \frac{s^2}{2}\cdot\frac{\mu_3 e^{\lambda(t-s)}}{\lambda/\alpha+\mu_3 e^{\lambda(t-s)}/\lambda}\,ds\right)\\ &= \exp\left(-\gamma\mu_1\frac{\mu_2}{\alpha}\int_0^t \frac{s^2}{2}\cdot\frac{(\alpha/\lambda)\mu_3 e^{\lambda(t-s)}}{1+(\alpha/\lambda)(\mu_3/\lambda)e^{\lambda(t-s)}}\,ds\right)\end{aligned}$$

Integrating by parts with $f(s) = s^2/2$ and $g'(s)$ is the fraction inside the integral

$$H_4(t) = \exp\left\{-\gamma\mu_1\frac{\mu_2}{\alpha}\int_0^t s\log(1+(\alpha/\lambda)(\mu_2/\lambda)\theta e^{\lambda(t-s)})\,ds\right\} \tag{47}$$

since $f(s)g(s) = 0$ when $s = 0$ and $s = t$. Changing variables $u = t - s$ gives the formula for $H_4(t)$. □

9 Conclusions

Here we have taken a probabilistic approach to analyze multi-stage models of cancer incidence. This leads to an intuitive proof of a simple and general formula for the distribution of the waiting time T_k for the first type k to appear

$$H_k(t) \equiv P(T_k \geq t) = \int_0^t \eta_{k-1}(s)\frac{\lambda}{\alpha}\cdot(1-G(t-s))\,ds \tag{48}$$

where $\eta_{k-1}(s) = N\mu_0\mu_1\cdots\mu_{k-1}s^{k-2}/(k-2)!$ is the rate type $k-1$ mutations are produced at time s, λ/α is the probability a type $k-1$ is successful, i.e., does not die out and

$$1-G(t-s) = \frac{\lambda/\alpha}{\lambda/\alpha+\mu e^{\lambda(t-s)}/\lambda}$$

is the probability a successful type $k-1$ born at time s produces a malignant cell by time t.

Differentiating (48) we can get a formula for the hazard rate $h_k(t) = -H_k'(t)/H_k(t)$. To do this it is convenient change variables $u = t - s$ and write $\Gamma_{k-1} = N\mu_0\mu_1\cdots\mu_{k-1}$

$$H_k(t) = \int_0^t \Gamma_{k-1}\frac{(t-u)^{k-2}}{(k-2)!}\frac{\lambda}{\alpha}(1-G(u))\,du$$

In the case $k = 2$ we have $(t-u)^{k-2}/(k-2)! \equiv 1$ so there is no t in the integrand and the derivative is

$$h_2(t) = N\mu_0 \frac{\lambda}{\alpha}(1 - G(t))$$

When $k \geq 3$ we have a positive power of $t - u$ so differentiating the upper limit does not contribute and the derivative is

$$h_k(t) = \int_0^t \Gamma_{k-1} \frac{(t-u)^{k-3}}{(k-3)!} \frac{\lambda}{\alpha}(1 - G(u)\, du$$

We have verified that our new formulas are almost exactly the same as the traditional ones for the k-stage model.

In Sections 4, 5, and 6 we considered four concrete applications that have been analyzed in the literature. In the case of pancreatic cancer, three and four stage models gave similar fits. See Figure 3. In the case of colon cancer, one gets almost identical fits from k-stage models with $k = 2, 3, 4, 5$. The parameter values for those fits (see Table 3) indicate how this is possible. The two stage fits have very small mutation rates while in the four and five stage fits, one or two mutation rates take large values. The pancreatic and colon cancer examples suggests that fitting k-stage models has little power to estimate the number of stages, but that power might be restored by constraining the parameter values to take on "realistic" values.

References

[1] Armitage, P. (1953) A note on the time-homogeneneous birth process. *J. Royal Statistical Society, B.* 15, 90

[2] Armitage, P., and Doll, R. (1954) The age distribution of cancer and a multi-stage theory of carcinogenesis. *British J. Cancer.* 8, 1–12

[3] Armitage, P., and Doll, R. (1957) A two-stage theory of carcinogenesis in relation to the age-distribution of human cancers. *British Journal of Cancer.* 11, 161–169

[4] Bailey, P. et al. (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature.* 531, 47–65

[5] Durrett, R. (2012) *Essentials of Stochastic Processes.* Springer, New York

[6] Durrett, R. (2015) *Branching Process Models of Cancer.* Springer, New York

[7] Durrett, R., and Moseley, S. (2010) Evolution of resistance and progression to disease during clonal expansion of cancer. *Theor. Pop. Biol.* 77, 42–48

[8] Fearon, E.R., and Vogelstein, B. (1990) A genetic model for colorectal tumorigenesis. *Cell.* 759–767

[9] Frank, S.A. (2007) *Dynamics of Cancer: Incidence, Inheritance and Evolution.* Princeton U. Press

[10] Fisher, J.C., and Hollomon, J.H. (1951) A hypothesis for the origin of cancer foci. *Cancer.* 4, 916–918

[11] Garraway, L.A., and Lander, E.S. (2013) Lessons from the cancer genome. *Cell.* 153, 17–37

[12] Heidenreich, W.F., Luebeck, E.G., and Moolgavkar, S.H. (1997) Some properties of the hazard function of the two-mutation clonal expansion model. *Risk Analysis.* 17 (1997), 391–399

[13] Hoogenveen, R.T., Clewell, H.J., Andersen, M.E., and Slob, W. (1999) An alternative exact solution of the two-stage clonal growth model of cancer. *Risk Analysis.* 19, 9–14

[14] Knudson, A.G. (1971) Mutation and cancer: Statistical study of retinoblastoma. *Proc. Natl. Acad. Sci.* 68, 820–823

[15] Knudson, A.G. (2001) Two genetic hits (more or less) to cancer. *Nature Reviews Cancer.* 1, 157–162

[16] Leubeck, E.G and Moolgavkar, S.H. (2002) Multistage carcinogenesis and the incidence of colorectal cancer. *Proc. Natl. Acad. Sci.* 99, 15095–15100

[17] Moolgavkar, S.H., Meza, R., and Turim, J. (2009) Pleural and peritoneal mesothelioma in SEER: age effects and termporal trends, 1973–2005. *Cancer Causes Control.* 20, 935–944

[18] Meza, R., and Chang, J.T. (2015) Multistage carcinogenesis and gthe incidence of thryoid cancer in the US by sex, race, stage and histology. *BMC Public Health.* 15, paper 789

[19] Meza, R., Jeon, J., Moolgavkar, S.H., and Luebeck, E.G. (2008) Age-specific incidence of cancer: Phases, transitions, and biological implications. *Proc. Natl. Acad. Sci.* 105, 16284–16289

[20] Nordling, C.O. (1953) A new theory on cancer inducing mechanism. *Brit. J. Cancer.* 7, 68–72

[21] The Cancer Genome Atlas Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* 487, 330-337

[22] The Cancer Genome Atlas Research Network. (2014) Integrated genomic characterization of pappilary thyroid carcinoma. *Cell.* 159–690

[23] Tomasetti, C., marchioni, L., nowak, M.A., parmigiani, G., and Vogelstein, B. (2015) Only three driver gene mutations are required for the developmetn of lung and colorectal cancers. *Proc. Natl. Acad. Sci.* 112, 118–123

[24] Vogelstein, B., and Kinzler, K.W. (1998) *The Genetic Basis of Human Cancer.* McGraw Hill

[25] Vogelstein, B, et al. (2013) Cancer genome landscapes. Science 339, 1546–1558