Multi-stage models for the failure of complex systems, cascading disasters, and the onset of disease

Anthony J. Webster

Nuffield Department of Population Health, Richard Doll Building, University of Oxford, Old Road Campus, Oxford, OX3 7LF, UK.

anthony.webster@ndph.ox.ac.uk

Abstract

Complex systems can fail through different routes, often progressing through a series of (rate-limiting) steps and modified by environmental exposures. The onset of disease, cancer in particular, is no different. Multi-stage models provide a simple but very general mathematical framework for studying the failure of complex systems, or equivalently, the onset of disease. They include the Armitage-Doll multi-stage cancer model as a particular case, and have potential to provide new insights into how failures and disease, arise and progress. A method described by E.T. Jaynes is developed to provide an analytical solution for a large class of these models, and highlights connections between the convolution of Laplace transforms, sums of random variables, and Schwinger/Feynmann parameterisations. Examples include: exact solutions to the Armitage-Doll model, the sum of Gamma-distributed variables with integer-valued shape parameters, a clonal-growth cancer model, and a model for cascading disasters. Applications and limitations of the approach are discussed in the context of recent cancer research. The model is sufficiently general to be used in many contexts, such as engineering, project management, disease progression, and disaster risk for example, allowing the estimation of failure rates in complex systems and projects. The intended result is a mathematical toolkit for applying multi-stage models to the study of failure rates in complex systems and to the onset of disease, cancer in particular.

1 Introduction

Complex systems such as a car can fail through many different routes, often requiring a sequence or combination of events for a component to fail. The same can be true for human disease, cancer in particular [1, 2, 3]. For example, cancer can arise through a sequence of steps such as genetic mutations, each of which must occur prior to cancer [4, 5, 6, 7, 8]. The considerable genetic variation between otherwise similar cancers [9, 10], suggests that similar cancers might arise through a variety of different paths.

Multi-stage models describe how systems can fail through one or more possible routes. They are sometimes described as "multi-step" or "multi-hit" models [11, 12], because each route typically requires failure of one or more sequential or non-sequential steps. Here we show that the model is easy to conceptualise and derive, and that many specific examples have analytical solutions or approximations, making it ideally suited to the construction of biologically- or physically-motivated models for the incidence of events such as diseases, disasters, or mechanical failures. A method described by E.T. Jaynes [13] generalises to give an exact analytical formula for the sums of random variables needed to evaluate the sequential model. This is evaluated for specific cases. Moolgavkar's exact solution [14] to the Armitage-Doll multistage cancer model is one example that is derived surprisingly easily, and is easily modified. The approach described here can incorporate simple models for a clonal expansion prior to cancer

detection [5, 6, 7], but as discussed in Sections 8 and 9, it may not be able to describe evolutionary competition or cancer-evolution in a changing micro-environment without additional modification. More generally, it is hoped that the mathematical framework can be used in a broad range of applications, including the modelling of other diseases [15, 16, 17, 18]. One example we briefly describe in Section 8 is modelling of "cascading disasters" [19], where each disaster can substantially modify the risk of subsequent (possibly different) disasters.

Conventional notation is used [20], with: probability densities $f(t)$, cumulative probability distributions $F(t) = \int_0^t f(t)$, a survival function $S(t) = 1 - F(t)$, hazard function $h(t) = f(t)/S(t)$, and cumulative hazard function $H(t) = \int_0^t h(y)dy$. Noting that $f(t) = -dS/dt$, it is easily seen that $H(t) = \int_0^t f(y)/S(y)dy = -\log S(t)$, $h(t) = -d\log S(t)/dt$, and $S(t) = \exp(-\int_0^t h(y)dy)$.

2 Failure by multiple possible routes

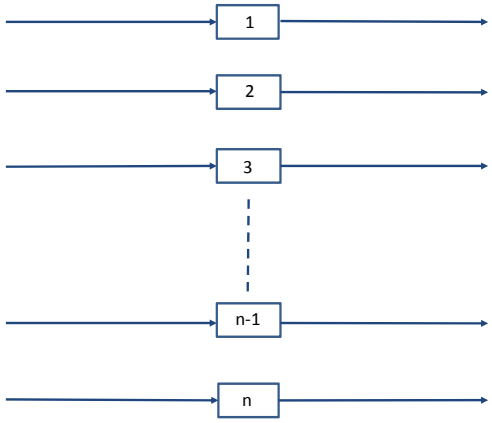

Figure 1: *In a complex system, failure can occur through many different routes (Eq. 1).*

Imagine that we can enumerate all possible routes 1 to n by which a failure can occur (Figure 1). The probability of surviving the ith of these routes after time t is $S_i(t)$, and consequently the probability of surviving all of these possible routes to failure $S(t)$ is,

$$S(t) = \Pi_{i=1}^{n} S_i(t) \tag{1}$$

or in terms of cumulative hazard functions with $S_i(t) = e^{-H_i(t)}$,

$$S(t) = \exp\left\{-\sum_{i=1}^{n} H_i(t)\right\} \tag{2}$$

The system's hazard rate for failure by any of the routes is,

$$\begin{aligned} h(t) &= -\tfrac{d}{dt}\log\left(S(t)\right) \\ &= -\textstyle\sum_{i=1}^{n} \tfrac{d}{dt}\log\left(S_i(t)\right) \\ &= \textstyle\sum_{i=1}^{n} h_i(t) \end{aligned} \tag{3}$$

and $H(t) = \sum_{i=1}^{n} H_i(t)$. In words, if failure can occur by any of n possible routes, the overall hazard of failure equals the sum of the hazard of failure by all the individual routes.

A few notes on Eq. 2 and its application to cancer modelling. Firstly, if the sth route to failure is much more likely than the others, with $H_s \gg H_j$ for $s \neq j$, then $S(t) = \exp\left\{-H_s(t) + \left(1 + O\left(\sum_{i \neq s} H_i/H_s\right)\right)\right\} \simeq \exp\{-H_s(t)\}$, which could represent the most likely sequence of mutations in a cancer model for example. Due to different manufacturing processes, genetic backgrounds, chance processes or exposures (e.g. prior to adulthood), this most probable route to failure could differ between individuals. Secondly, the stem cell cancer model assumes that cancer can occur through any of n_s equivalent stem cells in a tissue, for which Eq. 2 is modified to, $S = \exp\{-n_s \sum_{i=1}^{n} H_i(t)\}$. So a greater number of stem cells is expected to increase cancer risk, as is observed [21, 22]. Thirdly, most cancers are sufficiently rare that $S \sim 1$. As a consequence, many cancer models (implicity or explicitly) assume $S \simeq 1 - n_s \sum_{i=1}^{n} H_i(t)$ and $f = -dS/dt \simeq n_s \sum_{i=1}^{n} h_i(t)$, a limit emphasised in the Appendix of Moolgavkar [14].

3 Failure requiring m independent events

Figure 2: *Failure by the ith path at time t requires m_i independent failures to occur in any order, with the last failure at time t (Eq. 5).*

Often failure by a particular path will require more than one failure to occur independently. Consider firstly when there are m_i steps to failure, and the order of failure is unimportant (Figure 2). The probability of surviving failure by the ith route, $S_i(t)$ is,

$$\begin{aligned} S_i(t) &= P\,(\text{survive any one or more, necessary step for failure}) \\ &= 1 - P\,(\text{survive all the steps}) \\ &= 1 - \Pi_{j=1}^{m_i} F_{ij}(t) \end{aligned} \tag{4}$$

where $F_{ij}(t)$ is the cumulative probability distribution for failure of the jth step on the ith route within time t. Writing $S_{ij}(t) = 1 - F_{ij}(t)$, this can alternately be written as,

$$S_i(t) = 1 - \Pi_{j=1}^{m_i} \left(1 - S_{ij}(t)\right) \tag{5}$$

4 Relation to recent multi-stage cancer models

It may be helpful to explain how Eqs. 1 and 4 are used in recently described multi-stage cancer models [23, 24, 25]. If we take a rate of mutations μ_j per cell division for each of the rate-limiting mutational steps 1 to j, and d_i divisions of cell i, then the probability of a stem cell surviving without the jth rate limiting mutation is $S_{ij} = (1 - \mu_j)^{d_i}$. Similarly, the probability of a given stem cell having mutation j is $F_{ij} = 1 - (1 - \mu_j)^{d_i}$. This is the solution of Zhang et al. [24] to the recursive formula of Wu et al. [23] (see Appendix of Zhang et al. [24] for details). Using Eq. 4, the survival of the ith stem cell is described by,

$$S_i = 1 - \Pi_{j=1}^{m_i} \left(1 - (1 - \mu_j)^{d_i}\right) \tag{6}$$

Now assuming all n stem cells are equivalent and have equal rates $\mu_i = \mu_j$ for all i, j, and consider only one path to cancer with m mutational steps, then,

$$S_i = 1 - \left(1 - (1 - \mu)^{d}\right)^{m} \tag{7}$$

and,

$$\begin{aligned} S &= \Pi_{i=1}^{n} S_i \\ &= \left(1 - \left(1 - (1-\mu)^d\right)^m\right)^n \end{aligned} \tag{8}$$

The probability of cancer within m divisions, often referred to as "theoretical lifetime intrinsic cancer risk", is,

$$F = 1 - \left(1 - \left(1 - (1-\mu)^d\right)^m\right)^n \tag{9}$$

This is the equation derived by Calabrese and Shibata [25], and that Zhang found as the solution to the model of Wu et al [24, 23].

Therefore, in addition to the models of Wu and Calabrese being equivalent cancer models needing m mutational steps, the models also assume that the order of the steps is not important. This differs from the original Armitage-Doll model that considered a sequential set of rate-limiting steps, and was exactly solved by Moolgavkar [14]. Eqs. 8 and 9 are equivalent to assuming: (i) equivalent stem cells, (ii) a single path to cancer, (iii) equivalent divisions per stem cell, and, (iv) equivalent mutation rates for all steps.

Despite the differences in modelling assumptions for Eq. 9 and the Armitage-Doll model, their predictions can be quantitatively similar. To see this, use the Armitage-Doll approximation of $\mu d \ll 1$, to expand,

$$(1-\mu)^d = \exp\left(d \log(1-\mu)\right) \simeq \exp\left(\mu d\right) \tag{10}$$

If cell divisions are approximately uniform in time, then we can replace μd with μt, with μ now a rate per unit time. Then expanding $\exp(-\mu t) \simeq 1 - \mu t$, gives,

$$F = 1 - \left(1 - \left(1 - (1-\mu)^d\right)^m\right)^{n_s} \simeq 1 - \left(1 - (\mu t)^m\right)^{n_s} \simeq n_s \left(\mu t\right)^m \tag{11}$$

The incidence rate $h = f/S$ is then $h \simeq n_s \mu^m t^{m-1}$, the same as the original (approximate) Armitage-Doll solution [2]. This approximate solution is expected to become inaccurate at sufficiently long times.

An equivalent expression to Eq. 8 was known to Armitage, Doll, and Pike since at least 1965 [26], as was its limiting behaviour for large n. The authors [26] emphasised that many different forms for the $F_i(t_i)$ could produce approximately the same observed $F(t)$, especially for large n, with the behaviour of $F(t)$ being dominated by the small t behaviour of $F_i(t)$. As a result, for sufficiently small times power-law behaviour for $F(t)$ is likely, and if longer times were observable then an extreme value distribution would be expected [26, 27, 4]. However the power-law approximation can fail for important cases with extra rate-limiting steps such as a clonal expansion [5, 6, 7]. It seems likely that a model that includes clonal expansion and cancer detection is needed for cancer modelling, but the power law approximation could be used for all but the penultimate step, for example. A general methodology that includes this approach is described next, and examples are given in the subsequent section 6. The results and examples of sections 5 and 6 are intended to have a broad range of applications.

5 Failure requiring m sequential steps

Some failures require a *sequence* of independent events to occur, each following the one before (Figure 3). A well-known example is the Armitage-Doll multistage cancer model, that requires a sequence of m mutations (failures), that each occur with a different constant rate. The probability density for failure time is the pdf for a sum of the m independent times t_j to failure at each step in the sequence, each of which may have a different probability density function $f_j(t_j)$. A general method for evaluating the

Figure 3: *Failure by the ith path at time t requires an ordered sequence of failures, with the last failure at time t (Eqs. 16 and 18).*

probability density is outlined below, adapting a method described by Jaynes [13] (page 569).

Take $T_i \sim f_i(t_i)$ as random variables. Then use marginalisation to write $P\left(\sum_{j=1}^{m} T_j = t\right)$ in terms of $P\left(\sum_{j=1}^{m} T_j = t, T_1 = t_1, \ldots, T_m = t_m\right)$, where (A, B, C) is read as "A and B and C", and expand using the product rule $P(A, B) = P(A|B)P(B)$,

$$\begin{array}{lll} P\left(\sum_{j=1}^{m} T_j = t\right) & = \int_0^\infty dt_1 \cdots \int_0^\infty dt_m & P\left(\sum_{j=1}^{m} T_j = t, T_1 = t_1, \ldots, T_m = t_m\right) \\ & = \int_0^\infty dt_1 \cdots \int_0^\infty dt_m & P\left(\sum_{j=1}^{m} T_j = t | T_1 = t_1, \ldots, T_m = t_m\right) \\ & & \times P\left(T_1 = t_1, \ldots, T_m = t_m\right) \end{array} \tag{12}$$

Noting that $P\left(\sum_{j=1}^{m} T_j = t | T_1 = t_1, \ldots, T_m = t_m\right)$ is zero for $t \neq \sum_{j=1}^{m} t_j$ and $1 = \int_0^\infty dt P\left(\sum_{j=1}^{m} T_j = t | T_1 = t_1, \ldots, T_m = t_m\right)$, indicates that it is identical to a Dirac delta function $\delta\left(t - \sum_{j=1}^{m} t_j\right)$. For independent events $P(T_1 = t_1, \ldots, T_m = t_m) = \Pi_{j=1}^{m} f_j(t_j)$ where $f_j(t_j) \equiv P_j(T_j = t_j)$. Writing $f(t) \equiv P\left(\sum_{j=1}^{m} T_j = t\right)$, then gives,

$$f(t) = \int_0^\infty dt_1 \cdots \int_0^\infty dt_m \Pi_{j=1}^{m} f_j(t_j) \delta\left(t - \sum_{j=1}^{m} t_j\right) \tag{13}$$

To evaluate the integrals, take the Laplace transform with respect to t, to give,

$$\mathcal{L}[f] = \int_0^\infty e^{-st} f(t) dt = \int_0^\infty dt_1 \cdots \int_0^\infty dt_m \Pi_{j=1}^{m} f_j\left(t_j\right) e^{-s(t_1 + \cdots + t_m)} \tag{14}$$

This factorises as,

$$\mathcal{L}[f] = \Pi_{j=1}^{m} \int_0^\infty dt_j f_j(t_j) e^{-st_j} \tag{15}$$

Giving a general analytical solution as,

$$f(t) = \mathcal{L}^{-1}\left\{\Pi_{j=1}^{m} \mathcal{L}\left[f_j(t_j)\right]\right\} \tag{16}$$

where $\mathcal{L}^{-1}$ is the inverse Laplace transform, and $\mathcal{L}\left[f_j(t_j)\right] = \int_0^\infty dt_j f_j(t_j) e^{-st_j}$ with the same variable s for each value of j. Eq. 15 is similar to the relationship between moment generating functions $M_i(s) = \sum_{t_i=0}^{\infty} e^{st_i} p_i(t_i)$ of discrete probability distributions $p_i(t_i)$, and the moment generating function $M(s)$ for $t = \sum_{i=1}^{m} t_i$, that has,

$$M(s) = \Pi_{i=1}^{m} M_i(s) \tag{17}$$

whose derivation is analogous to Eq. 17 but with integrals replaced by sums. The survival and hazard functions for $f(t)$ can be obtained from Eq. 16 in the usual way. For example,

$$\begin{array}{ll} S_i(t) & = \int_t^\infty f_i(y) dy \\ & = \int_t^\infty \mathcal{L}^{-1}\left\{\Pi_{j=1}^{m_i} \mathcal{L}\left[f_{ij}(t_{ij})\right]\right\} dy \end{array} \tag{18}$$

that can be used in combination with Eq. 1. A number of valuable results are easy to evaluate using Eq. 16, as is illustrated in the next section.

A useful related result is,

$$f(t) = \mathcal{L}^{-1}\left\{\mathcal{L}\left[f\left(\sum_{j=1}^{n-1} t_j\right)\right]\mathcal{L}\left[f_n(t_n)\right]\right\} \tag{19}$$

that can be inferred from Eq. 16 with $m = 2$,

$$f\left(t = t_1 + t_2\right) = \mathcal{L}^{-1}\left\{\mathcal{L}\left[f_1(t_1)\right]\mathcal{L}\left[f_2(t_2)\right]\right\} \tag{20}$$

by replacing $f_1(t_1)$ with $f(\sum_{j=1}^{n-1} t_j)$ and $f_2(t_2)$ with $f_n(t_n)$. Eq. 20 can be solved using the convolution theorem for Laplace transforms, that gives,

$$f\left(t = t_1 + t_2\right) = \int_0^t f_1(\tau) f_2(t-\tau) d\tau \tag{21}$$

which is sometimes easier to evaluate than two Laplace transforms and their inverse. In general, solutions can be presented in terms of multiple convolutions if it is preferable to do so. Eqs. 19 and 21 are particularly useful for combining a known solution for the sum of $(n-1)$ samples such as for cancer initiation, with a differently distributed nth sample, such as the waiting time to detect a growing cancer. A final remark applies to the sum of random variables whose domain extends from $-\infty$ to ∞, as opposed to the range 0 to ∞ considered so far. In that case an analogous calculation using a Fourier transform with respect to t in Eq. 13 leads to analogous results in terms of Fourier transforms, with $\mathcal{F}[f_i(t_i)] = \int_{-\infty}^{\infty} f_i(t_i) e^{imt_i} dt_i$ in place of Laplace transforms, resulting in,

$$f(t) = \mathcal{F}^{-1}\left\{\Pi_{j=1}^{m} \mathcal{F}\left[f_j(t_j)\right]\right\} \tag{22}$$

Eq. 22 is mentioned for completeness, but is not used here.

A general solution to Eq. 16 can be given in terms of definite integrals, with,

$$\begin{aligned} f(t) \quad &= \mathcal{L}^{-1}\left\{\Pi_{j=1}^{m} \mathcal{L}\left[f_j(t_j)\right]\right\} \\ &= t^{m-1} \int_0^1 dy_1 \cdots \int_0^1 dy_{m-1} y_1^0 y_2^1 \ldots y_{m-1}^{n_m-1} \\ &\quad f_1(t y_1 \ldots y_{m-1}) f_2(t(1-y_1) y_2 \ldots y_{m-1}) f_3(t(1-y_2) y_3 \ldots y_{m-1}) \ldots \\ &\quad f_{m-1}(t(1-y_{m-2}) y_{m-1}) f_m(t(1-y_{m-1})) \end{aligned} \tag{23}$$

This can sometimes be easier to evaluate or approximate than Eq. 16. A derivation is given in Appendix 62. Eq. 23 allows a generalised Schwinger/Feynmann parameterisation [28] to be derived. Writing $g_j(s) = \mathcal{L}\left[f_j(t_j)\right]$ and taking the Laplace transform of both sides of Eq. 23, gives,

$$\begin{aligned} \Pi_{j=1}^{m} g_j(s) \quad &= \int_0^1 dy_1 \cdots \int_0^1 dy_{m-1} y_1^0 y_2^1 \ldots y_{m-1}^{n_m-1} \\ &\quad \mathcal{L}\left[t^{m-1} \mathcal{L}^{-1}\left\{g_1(s)\right\}(t y_1 \ldots y_{m-1}) \mathcal{L}^{-1}\left\{g_2(s)\right\}(t(1-y_1) y_2 \ldots y_{m-1}) \ldots \right. \\ &\quad \left. \mathcal{L}^{-1}\left\{g_m(s)\right\}(t(1-y_{m-1}))\right] \end{aligned} \tag{24}$$

which includes some well known Schwinger/Feynmann parameterisations as special cases. This is discussed further in the Appendix.

6 Modelling sequential events - examples

In the following examples we consider the time $t = \sum_{i=1}^{m} t_i$ for a sequence of events, with possibly different distributions $f_i(t_i)$ for the time between the $(i-1)$th and ith

event. Some of the results are well-known but not usually presented this way, others are new or poorly known. We will use the Laplace transforms (and their inverses), of,

$$\mathcal{L}^{-1}\mathcal{L}[t^p] = \mathcal{L}^{-1}[\Gamma(p+1)/s^{p+1}] = t^p \tag{25}$$

and,

$$\mathcal{L}^{-1}\mathcal{L}[t^p e^{-\mu t}] = \mathcal{L}^{-1}[\Gamma(p+1)/(s+\mu)^{p+1}] = t^p e^{-\mu t} \tag{26}$$

Sums of gamma distributed samples (equal rates): Using Eq. 16, the sum of m gamma distributed variables with equal rate parameters μ, and $f_i(t_i) = \mu^{p_i} t_i^{p_i-1} e^{-\mu t_i}/\Gamma(p_i)$, are distributed as,

$$\begin{aligned} f(t) = \mathcal{L}^{-1}\left\{\Pi_{i=1}^m \mathcal{L}\left[\mu^{p_i}\frac{t_i^{p_i-1}e^{-\mu t_i}}{\Gamma(p_i)]}\right]\right\} &= \mathcal{L}^{-1}\left\{\Pi_{i=1}^m \frac{\mu^{p_i}}{(s+\mu)^{p_i}}\right\} \\ &= \mathcal{L}^{-1}\left\{\frac{\mu^{\sum_{i=1}^m p_i}}{(s+\mu)^{\sum_{i=1}^m p_i}}\right\} \\ &= \mu^{\sum_{i=1}^m p_i}\frac{t^{\sum_{i=1}^m p_i} e^{-\mu t}}{\Gamma\left(\sum_{i=1}^m p_i\right)} \end{aligned} \tag{27}$$

For a sum of m exponentially distributed variables with $\{p_i = 1\}$, this simplifies to $f(t) = \mu^m t^{m-1} e^{-\mu t}/\Gamma(m)$, a Gamma distribution.

Power law approximations: For many situations such as most diseases, you are unlikely to get any particular disease during your lifetime. In those cases the probability of survival over a lifetime is close to 1, and the probability density function $f_i = h_i/S_i$, can be approximated by $f_i \simeq h_i$, that in turn can often be approximated by a power of time with $f_i \simeq h_i \simeq \mu_i t_i^{p_i}$. Then we have,

$$\begin{aligned} f(t) = \mathcal{L}^{-1}\left\{\Pi_{i=1}^m \mathcal{L}\left[\mu_i t_i^{p_i}\right]\right\} &= \mathcal{L}^{-1}\left\{\Pi_{i=1}^m \frac{\mu_i \Gamma(1+p_i)}{s^{1+p_i}}\right\} \\ &= \Pi_{i=1}^m \left(\mu_i \Gamma(1+p_i)\right) \mathcal{L}^{-1}\left\{\frac{1}{s^{m+\sum_{i=1}^m p_i}}\right\} \\ &= \Pi_{i=1}^m \left(\mu_i \Gamma(1+p_i)\right) \frac{t^{-1+m+\sum_{i=1}^m p_i}}{\Gamma(m+\sum_{i=1}^m p_i)} \end{aligned} \tag{28}$$

The Armitage-Doll model: A well known example of this approximation Eq. 28, is (implicitly) in the original approximate solution to the Armitage-Doll multi-stage cancer model. Taking a constant hazard at each step, and approximating $f_i \simeq h_i = \mu_i$, then Eq. 28 gives,

$$f(t) = \mathcal{L}^{-1}\left\{\Pi_{i=1}^m \mathcal{L}\left[\mu_i\right]\right\} = \left[\Pi_{i=1}^m \mu_i\right] \frac{t^{m-1}}{\Gamma(m)} \tag{29}$$

as was used in the original Armitage-Doll paper. Note that an equivalent time-dependence can be produced by a different combination of hazard functions with $h_i \sim t_i^{p_i}$ and $\tilde{m}$ steps, provided $m = \tilde{m} + \sum_{i=1}^{\tilde{m}} p_i$. For example, if $m = 6$, there could be 3 steps with $p = 1$, or 2 steps with $p = 2$, or 3 steps with $p_1 = 0$, $p_2 = 1$, and $p_3 = 2$, or some more complex combination. If the full pdfs are modelled at each step as opposed to their polynomial approximation, then this flexibility is reduced, as is the case for Moolgavkar's exact solution to the Armitage-Doll model that is described next.

Moolgavkar's exact solution to the Armitage-Doll model: Moolgavkar's exact solution to the Armitage-Doll model is the solution of,

$$f(t) = \mathcal{L}^{-1}\left\{\Pi_{i=1}^m \mathcal{L}\left[\mu_i e^{-\mu_i t_i}\right]\right\} = \mathcal{L}^{-1}\left\{\Pi_{i=1}^m \frac{\mu_i}{s+\mu_i}\right] \tag{30}$$

For example, if $n = 3$ then,

$$\mathcal{L}^{-1}\left\{\Pi_{i=1}^3 \mathcal{L}\left[\mu_i e^{-\mu_i t_i}\right]\right\} = \mu_1\mu_2\mu_3 \mathcal{L}^{-1}\left\{\frac{1}{(s+\mu_1)}\frac{1}{(s+\mu_2)}\frac{1}{(s+\mu_3)}\right\} \tag{31}$$

Using partial fractions, we can write,

$$\frac{1}{s+\mu_1}\frac{1}{s+\mu_2}\frac{1}{s+\mu_3} = \frac{1}{s+\mu_1}\frac{1}{(\mu_1-\mu_2)(\mu_1-\mu_3)} + \frac{1}{s+\mu_2}\frac{1}{(\mu_2-\mu_1)(\mu_2-\mu_3)} + \frac{1}{s+\mu_3}\frac{1}{(\mu_3-\mu_1)(\mu_3-\mu_2)} \tag{32}$$

Allowing the inverse Laplace transforms to be easily evaluated, giving,

$$\begin{aligned} f(t) &= \mathcal{L}^{-1}\left\{\Pi_{i=1}^{3}\mathcal{L}\left[\mu_i e^{-\mu_i t}\right]\right\} \\ &= \mu_1\mu_2\mu_3\left[\frac{e^{-\mu_1 t}}{(\mu_1-\mu_2)(\mu_1-\mu_3)} + \frac{e^{-\mu_2 t}}{(\mu_2-\mu_1)(\mu_2-\mu_3)} + \frac{e^{-\mu_3 t}}{(\mu_3-\mu_1)(\mu_3-\mu_2)}\right] \end{aligned} \tag{33}$$

Note that the result is independent of the order of sequential events, but unlike the approximate solution to the Armitage Doll model [2], the exact solution allows less variability in the underlying models that can produce it. Also note that the leading order terms of an expansion in t cancel exactly, to give identical leading-order behaviour as for a power-law approximation (with $p = 0$)

A general solution can be formed using a Schwinger/Feynmann parameterisation [28] of,

$$\Pi_{i=1}^{m}\frac{1}{\mu_i} = \Gamma(m)\int_0^1 dy_1\int_0^{y_1} dy_2\cdots\int_0^{y_{m-2}} dy_{m-1}\frac{1}{(\mu_1 y_{m-1}+\mu_2(y_{m-2}-y_{m-1})+\cdots+\mu_m(1-y_1))^m} \tag{34}$$

Replacing μ_i with $s+\mu_i$ in Eq. 34, then we can write Eq. 30 as,

$$\begin{aligned} &\mathcal{L}^{-1}\left\{\Pi_{i=1}^{m}\frac{\mu_i}{s+\mu_i}\right\} \\ &= \left(\Pi_{i=1}^{m}\mu_i\right)\Gamma(m)\times \\ &\int_0^1 dy_1\int_0^{y_1} dy_2\cdots\int_0^{y_{m-2}} dy_{m-1}\mathcal{L}^{-1}\left\{\frac{1}{(s+\mu_1 y_{m-1}+\mu_2(y_{m-2}-y_{m-1})+\cdots+\mu_m(1-y_1))^m}\right\} \\ &= \left(\Pi_{i=1}^{m}\mu_i\right)t^{m-1}\times \\ &\int_0^1 dy_1\int_0^{y_1} dy_2\cdots\int_0^{y_{m-2}} dy_{m-1}e^{-(\mu_1 y_{m-1}+\mu_2(y_{m-2}-y_{m-1})+\cdots+\mu_m(1-y_1))t} \end{aligned} \tag{35}$$

(which is simpler, but equivalent in effect, to repeatedly using the convolution formula). Completing the integrals will generate Moolgavkar's solution for a given value of m. For example, taking $m = 3$ and integrating once gives,

$$\mathcal{L}^{-1}\left\{\Pi_{i=1}^{3}\frac{\mu_i}{s+\mu_i}\right\} = \frac{te^{-\mu_3 t}}{(\mu_2-\mu_1)}\int_0^1 dx_1\left(e^{-x_1 t(\mu_1-\mu_3)} - e^{-x_1 t(\mu_2-\mu_3)}\right) \tag{36}$$

Integrating a second time, and simplifying, gives Eq. 33. The relationships between Schwinger/Feynmann parameterisations, Laplace transforms, and the convolution theorem are discussed further in the Appendix.

Moolgavkar [14] used induction to provide an explicit formula for $f(t)$, with,

$$f(t) = \left(\Pi_{i=1}^{m}\mu_i\right)\sum_{i=1}^{m}\chi_i(m)e^{-\mu_i t} \tag{37}$$

where,

$$\chi_i(m) = \frac{1}{(\mu_1-\mu_i)(\mu_2-\mu_i)\ldots(\mu_{i-1}-\mu_i)(\mu_{i+1}-\mu_i)\ldots(\mu_m-\mu_i)} \tag{38}$$

For small times the terms in a Taylor expansion of Eq. 37 cancel exactly, so that $f(t) \simeq \left(\Pi_{i=1}^{m}\mu_i\right)t^{m-1}$, as expected. This feature could be useful for approximating a normalised function when the early-time behaviour approximates an integer power of time. Further uses of Moolgavkar's solution are discussed next.

Sums of gamma distributed samples (with different rates): A useful mathematical result can be found by combining the Laplace transform of Moolgavkar's solution

Eq. 37 for $f\left(t=\sum_{i=1}^{m} t_i\right)$ with Eq. 30, to give an explicit formula for a partial fraction decomposition of the product $\Pi_{i=1}^{m} \frac{1}{s+\mu_i}$, as,

$$\Pi_{i=1}^{m} \frac{1}{s+\mu_i} = \sum_{i=1}^{m} \frac{\chi_i(m)}{s+\mu_i} \tag{39}$$

This can be useful in various contexts. For example, consider m Gamma distributions $f_i(t_i) = \mu_i^{p_i} t_i^{p_i-1} e^{-\mu_i t_i}/\Gamma(p_i)$ with different integer-valued shape parameters p_i, and $\mathcal{L}[f_i] = \mu_i^{p_i}/(s+\mu_i)^{p_i}$. Eq. 16 gives $f(t) = \left(\Pi_{i=1}^{m} \mu_i^{p_i}\right) \mathcal{L}^{-1}\left\{\Pi_{i=1}^{m} 1/(s+\mu_i)^{p_i}\right\}$, so firstly use the integer-valued property of $\{p_i\}$ to write,

$$\begin{aligned} \mathcal{L}^{-1}\left\{\Pi_{i=1}^{m} \tfrac{1}{(s+\mu_i)^{p_i}}\right\} &= \mathcal{L}^{-1}\left\{\Pi_{i=1}^{m} \tfrac{(-1)^{p_i-1}}{(p_i-1)!} \tfrac{\partial^{p_i-1}}{\partial \mu_i^{p_i-1}} \tfrac{1}{(s+\mu_i)}\right\} \\ &= \mathcal{L}^{-1}\left\{\Pi_{j=1}^{m} \tfrac{(-1)^{p_j-1}}{(p_j-1)!} \tfrac{\partial^{p_j-1}}{\partial \mu_j^{p_j-1}} \Pi_{i=1}^{m} \tfrac{1}{(s+\mu_i)}\right\} \end{aligned} \tag{40}$$

where the product of differential operators can be taken outside the product of Laplace transforms because $\partial/\partial\mu_i(1/(s+\mu_j))$ is zero for $i \neq j$. Using Eq. 39 we can replace the product of Laplace transforms with a sum, giving,

$$\mathcal{L}^{-1}\left\{\Pi_{i=1}^{m} \frac{1}{(s+\mu_i)^{p_i}}\right\} = \mathcal{L}^{-1}\left\{\Pi_{j=1}^{m} \frac{(-1)^{p_j-1}}{(p_j-1)!} \frac{\partial^{p_j-1}}{\partial \mu_j^{p_j-1}} \sum_{i=1}^{m} \frac{\chi_i(m)}{(s+\mu_i)}\right\} \tag{41}$$

The Laplace transform has now been simplified to a sum of terms in $1/(s+\mu_i)$, whose inverse Laplace transforms are easy to evaluate. Taking the inverse Laplace transform $\mathcal{L}^{-1}[1/(s+\mu_i)] = e^{-\mu_i t}$, and including the product $\Pi_{i=1}^{m} \mu_i^{p_i}$, gives,

$$f(t) = \left(\Pi_{i=1}^{m} \mu_i^{p_i}\right) \Pi_{j=1}^{m} \frac{(-1)^{p_j-1}}{(p_j-1)!} \frac{\partial^{p_j-1}}{\partial \mu_j^{p_j-1}} \sum_{i=1}^{m} \chi_i(m) e^{-\mu_i t} \tag{42}$$

as a general solution for sums of Gamma distributed samples with integer-valued shape parameters p_i (and arbitrary rate parameters μ_i). Eq. 42 is most easily evaluated with a symbolic algebra package.

If $p_i = p$ are equal, then Eq. 42 may be simplified further by writing,

$$f(t) = \left(\Pi_{i=1}^{m} \mu_i^{p}\right) \sum_{i=1}^{m} \frac{(-1)^{p-1}}{(p-1)!} \frac{\partial^{p-1}}{\partial \mu_i^{p-1}} \Pi_{j\neq i} \frac{(-1)^{p-1}}{(p-1)!} \frac{\partial^{p-1}}{\partial \mu_j^{p-1}} \left[\chi_i(m) e^{-\mu_i t}\right] \tag{43}$$

and noting that,

$$\Pi_{j\neq i} \frac{(-1)^{p-1}}{(p-1)!} \frac{\partial^{p-1}}{\partial \mu_j^{p-1}} \left[\chi_i(m) e^{-\mu_i t}\right] = \chi_i(m)^p e^{-\mu_i t} \tag{44}$$

because for $j \neq i$ there is exactly one factor $1/(\mu_j - \mu_i)$ in $\chi_i(m)$. This leaves,

$$f(t) = \left(\Pi_{i=1}^{m} \mu_i^{p}\right) \sum_{i=1}^{m} \frac{(-1)^{p-1}}{(p-1)!} \frac{\partial^{p-1}}{\partial \mu_i^{p-1}} \left[\chi_i(m)^p e^{-\mu_i t}\right] \tag{45}$$

for sums of Gamma distributed samples with the same integer-valued shape parameter p (and arbitrary rate parameters μ_i).

For example, if $p = 1$ then Eq. 45 becomes Moolgavkar's Eq. 37. Alternatively, if e.g. $p = 2$, then we have,

$$f(t) = \left(\Pi_{i=1}^{m} \mu_i^{2}\right) \sum_{i=1}^{m} \chi_i(m)^2 e^{-\mu_i t} \left[t - 2 \sum_{j\neq i} \frac{1}{(\mu_j - \mu_i)}\right] \tag{46}$$

for the sum of Gamma distributions with shape parameters $p = 2$ and arbitrary rate parameters, and $\chi_i(m)$ as defined in Eq. 38. If we also let e.g. $m = 2$, $\mu_2 = \mu_1 + \epsilon$, and $\epsilon \to 0$, then Eq. 46 tends to $\mu_1^4 t^3 e^{-\mu_1 t}/3!$, for the sum of two Gamma distributed variables with rate μ_1 and $p = 2$, in agreement with Eq. 27.

Sums of samples with different distributions: An advantage of the method described above, is that it is often easy to calculate pdfs for sums of differently distributed samples. For the first example, consider two samples from the same (or very similar) exponential distribution, and a third from a different exponential distribution. The result can be obtained by writing $\mu_3 = \mu_2 + \epsilon$ in Eq. 33, and letting $\epsilon \to 0$. Considering the terms involving exponents of μ_2 and μ_3,

$$\begin{aligned} \frac{e^{-\mu_2 t}}{(\mu_2-\mu_1)(\mu_2-\mu_3)} + \frac{e^{-\mu_3 t}}{(\mu_3-\mu_1)(\mu_3-\mu_2)} &= \frac{e^{-\mu_2 t}}{(\mu_2-\mu_1)\epsilon}\left(-1 + \frac{e^{-\epsilon t}}{1+\epsilon/(\mu_2-\mu_1)}\right) \\ &= \frac{e^{-\mu_2 t}}{\mu_2-\mu_1}\left(-1 + \left(1 - \epsilon t - \frac{\epsilon}{\mu_2-\mu_1} + O\left(\epsilon^2\right)\right)\right) \\ &= \left[-\frac{te^{-\mu_2 t}}{\mu_2-\mu_1} - \frac{e^{-\mu_2 t}}{(\mu_2-\mu_1)^2}\right](1 + O(\epsilon)) \end{aligned} \tag{47}$$

Using Eq. 33 and letting $\epsilon \to 0$, gives,

$$\begin{aligned} &\mu_1\mu_2\mu_3\left[\frac{e^{-\mu_1 t}}{(\mu_1-\mu_2)(\mu_1-\mu_3)} + \frac{e^{-\mu_2 t}}{(\mu_2-\mu_1)(\mu_2-\mu_3)} + \frac{e^{-\mu_3 t}}{(\mu_3-\mu_1)(\mu_3-\mu_2)}\right] \\ &\quad \to \mu_1{\mu_2}^2\left[\frac{e^{-\mu_1 t}-e^{-\mu_2 t}}{(\mu_1-\mu_2)^2} + \frac{te^{-\mu_2 t}}{(\mu_1-\mu_2)}\right] \end{aligned} \tag{48}$$

for the sum of three exponentially distributed variables, when exactly two have the same rate. Taking $\mu_2 = \mu_1 + \epsilon$ and letting $\epsilon \to 0$ in Eq. 48, gives a Gamma distribution $\mu_1^3 t^2 e^{-\mu_1 t}/2$, as it should for the sum of three exponentially distributed variables with equal rates (see Eq. 27 with $\{p_i = 0\}$). More generally, it can be seen that a sum of exponentially distributed samples with different rates, smoothly approximate a gamma distribution as the rates become increasingly similar, as expected from Eq. 27.

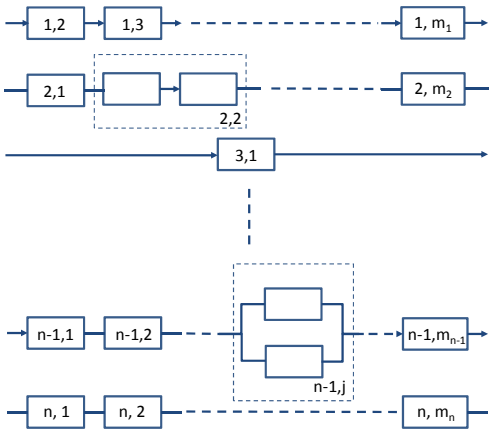

Figure 4: *Overall failure risk can be modelled as sequential steps (e.g. (1,1) to (1,m_1) using Eq. 5), and non-sequential steps (e.g. (n,1) to (n,m_n) using Eq. 16), that may be dependent on each other (e.g. Eq. 55). For the purposes of modelling, a sequence of dependent or multiple routes can be regarded as a single step (e.g. (2,2) or ($n-1$,j)).*

Failure involving a combination of sequential and non-sequential steps: If a path to failure involves a combination of sequential and non-sequential steps, then the necessary set of sequential steps can be considered as one of the non-sequential steps, with overall survival given by Eq. 1 and the survival for any sequential set of steps calculated from Eq. 18 (Figure 4).

7 Clonal-expansion cancer models

Clonal expansion is thought to be an essential element of cancer progression [29], and can modify the timing of cancer onset and detection [5, 6, 7, 32, 31, 30]. The growing number of cells at risk increases the probability of the next step in a sequence of mutations occurring, and if already cancerous, then it increases the likelihood of detection.

Some cancer models have a clonal expansion of cells as a rate-limiting step [5, 6, 7]. For example, Michor et al. [6] modelled clonal expansion of myeloid leukemia as logistic growth, with the likelihood of cancer detection (the hazard function), being proportional to the number of cancer cells. This gives a survival function for cancer detection of,

$$S_i(t) = \exp\left(-a\int_0^t x(y)dy\right) \tag{49}$$

where,

$$x(t) = \frac{1}{1+(N-1)e^{-ct}} \tag{50}$$

a, c, are rate constants, and N is the total number of cells prior to cancer initiation. Noting that $\int_0^t x(y)dy = \log(e^{ct}+(N-1))/c \to t$, as $t\to\infty$ and $x(t)\to 1$, then the tail of the survival curve falls exponentially towards zero with time.

Alternatively, we might expect the likelihood of cancer being diagnosed to continue to increase with time since the cancer is initiated. For example, a hazard function that is linear in time would give a Weibull distribution with $S(t) = e^{-at^2}$. It is unlikely that either this or the logistic model would be an equally good description for the detection of all cancers, although they may both be an improvement on a model without either. Both models have a single peak, but differ in the tail of their distribution, that falls as $\sim e^{-act}$ for the logistic model and $\sim e^{-at^2}$ for the Weibull model. Qualitatively, we might expect a delay between cancer initiation and the possibility of diagnosis, and diagnosis to occur almost inevitably within a reasonable time-period. Therefore a Weibull or Gamma distributed time to diagnosis may be reasonable for many cancers, with the shorter tail of the Weibull distribution making it more suitable approximation for cancers whose diagnosis is almost inevitable. (The possibility of misdiagnosis or death by another cause is not considered here.)

For example, noting that Moolgavkar's solution is a linear combination of exponential distributions, to combine it with a Weibull distribution for cancer detection $f_1(t_1) = -d/dt_1(e^{-bt_1^2/2})$, we can consider a single exponential term at a time. Taking $f_2(t_2) = ae^{-at_2}$, and using the convolution formula Eq. 21, we get,

$$\begin{aligned} f(t=t_1+t_2) &= \mathcal{L}^{-1}\left\{L[f_1(t_1)]L[f_2(t_2)]\right\} \\ &= a\int_0^t e^{-a(t-y)}\left(-\tfrac{d}{dy}e^{-by^2/2}\right)dy \\ &= a\left(e^{-at}-e^{-bt^2/2}\right) + a^2e^{-at}e^{a^2/2b}\int_0^t e^{-\frac{b}{2}\left(y-\frac{a}{b}\right)^2}dy \end{aligned} \tag{51}$$

where we integrated by parts to get the last line. This may be written as,

$$\begin{aligned} f(t) &= a\left(e^{-at}-e^{-bt^2/2}\right) \\ &\quad + a^2e^{-at}e^{a^2/2b}\sqrt{\tfrac{\pi}{2b}}\,\mathrm{erf}\left(\sqrt{\tfrac{b}{2}}\tfrac{a}{b}\right) \\ &\quad + a^2e^{-at}e^{a^2/2b}\sqrt{\tfrac{\pi}{2b}}\begin{cases} -\mathrm{erf}\left(\sqrt{\frac{b}{2}}\left(\frac{a}{b}-t\right)\right) & t<\frac{a}{b} \\ +\mathrm{erf}\left(\sqrt{\frac{b}{2}}\left(t-\frac{a}{b}\right)\right) & t\geq\frac{a}{b} \end{cases} \end{aligned} \tag{52}$$

with $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz$. Similarly for a Gamma distribution with $f_1 = b^p t^{p-1} e^{-bt}/\Gamma(p)$ and an exponential, $f_2(t_2) = ae^{-at}$, then assuming $b > a$,

$$\begin{aligned} f(t) &= \tfrac{b^p a}{\Gamma(p)} \int_0^t y^{p-1} e^{-by} e^{-a(t-y)} dy \\ &= b^p a \tfrac{e^{-at}}{(b-a)^p} \tfrac{1}{\Gamma(p)} \int_0^{t(b-a)} u^{p-1} e^{-u} du \\ &= b^p a \tfrac{e^{-at}}{(b-a)^p} \gamma(p, t(b-a)) \end{aligned} \tag{53}$$

where $\gamma(p, t(b-a))$ is the normalised lower incomplete Gamma function, which is available in most computational mathematics and statistics packages. If $a > b$ then f_1 and f_2 must be exchanged and the result is most easily evaluated numerically.

8 Cascading failures with dependent sequences of events

Now consider non-independent failures, where the failure of A changes the probability of a failure in B or C. In general, if the paths to failure are not independent of each other then the situation cannot be described by Eq. 1. For example [1], if step 1 of A prevents step 1 of B and vice-versa, then only one path can be followed. If the first step occurs at time t_1, the pdf for failure at time t is: $f(t) = S_A(t_1) f_B(t) + S_B(t_1) f_A(t)$, where $f_A(t)$ and $f_B(t)$ are the pdfs for path A and B if they were independent. This differs from Eq. 1 that has, $f(t) = -dS/dt = S_A(t) f_B(t) + S_B(t) f_A(t)$, that is independent of t_1. As a consequence, Eq. 1 may be inappropriate to describe phenomenon such as survival in the presence of natural selection, where competition for the same resource means that not all can survive. In some cases it may be possible to include a different model for the step or steps where Eq. 1 fails, analogously to the clonal expansion model [6] described in Section 6. But in principle, an alternative model may be required. We will return to this point in Section 9.

The rest of this section limits the discussion to situations where the paths to failure are independent, but where the failure-rate depends on the order of events. Important humanitarian examples are "cascading hazards" [19], where the risk of a disaster such as a mud slide is vastly increased if e.g. a wildfire occurs before it. An equivalent scenario would require m parts to fail for the system to fail, but the order in which the parts fail, modifies the probability of subsequent component failures. As an example, if three components A, B, and C, must fail, then we need to evaluate the probability of each of the 6 possible routes in turn, and obtain the overall failure probability from Eq. 1. Assuming the paths to failure are independent, then there are $m!$ routes, giving 6 in this example. Writing the 6 routes as, 1=ABC, 2=ACB, 3=BAC, 4=BCA, 5=CAB, 6=CBA, and reading e.g. ABC as "A, then B, then C", the survival probability is,

$$S(t) = \Pi_{i=1}^{6} S_i(t) \tag{54}$$

For failure by a particular route ABC we need the probability for the sequence of events, $A\&(\overline{B\&C})$, then $(B\&\bar{C})|A$, then $C|(AB)$. We can calculate this using Eq. 16, for example giving,

$$f_{ABC}(t) = \mathcal{L}^{-1} \left\{ \mathcal{L}\left[f_{A\&(\overline{B\&C})}(t_1) \right] \mathcal{L}\left[f_{(B\&\bar{C})|A}(t_2) \right] \mathcal{L}\left[f_{C|(AB)}(t_3) \right] \right\} \tag{55}$$

from which we can construct $S_1(t) = \int_t^\infty f_{ABC}(y) dy$.

Although in principle every term in e.g. Eqs. 54 and 55 need evaluating, there will be situations where results simplify. For example, if one route is much more probable than another - e.g. if it is approximately true that landslides only occur after deforestation,

[1] Thanks to Benjamin Cairns for this example.

that may be due to fire, then we only need to evaluate the probability distribution for that route. As another example, if all the f_i are exponentially distributed with different rates, then f_{ABC} will be described by Moolgavkar's solution. A more striking example is when there are very many potential routes to failure, as for the Armitage-Doll model where there are numerous stem cells that can cause cancer. In those cases, if the overall failure rate remains low, then the $f_i(t)$ in Eq. 55 must all be small with $S \simeq 1$ and $f \simeq h$, and can often be approximated by power laws. For that situation we have a general result that f_i, F_i, and H_i will be a powers of time, and Eq. 2 gives,

$$S(t) \simeq \exp\left\{-\sum_{i=1}^{n} a_i t^{p_i}\right\} \tag{56}$$

for some $a_i > 0$ and $p_i > 0$. Then $F(t) = 1 - S(t)$, $f(t) = -dS/dt$, and $h(t) \simeq f(t)$, can be approximated by a sum of power series in time. If one route is much more likely than the others then both $f(t)$ and $h(t)$ can be approximated as a single power of time, with the approximation best at early times, and a cross-over to different power-law behaviour at later times.

9 Cancer evolution, the tissue micro-environment, and model limitations

Cancer is increasingly viewed as an evolutionary process that is influenced by a combination of random and carcinogen-driven genetic and epigenetic changes [2, 3, 34, 33, 35, 36, 29, 21, 37], and an evolving tissue micro-environment [38, 39, 40, 41]. Although there is evidence that the number of stem cell divisions is more important for cancer risk than number of mutations [42, 43], the recognition that cells in a typical cancer are functionally and genetically diverse has helped explain cancers' resistance to treatment, and is suggesting alternative strategies to tackle the disease through either adaptive therapies [44, 45, 46, 47] or by modifying the tissue's micro-environment [48, 49, 39, 41]. This highlights two limitations of the multi-stage model described here.

Evolution: As noted in Section 8, Eq. 1 cannot necessarily model a competitive process such as natural selection, where the growth of one cancer variant can inhibit the growth of another. If the process can be described through a series of rate-limiting steps, then we could still approximate it with a form of Eq. 16. Otherwise, the time-dependence of a step with competitive evolutionary processes may need to be modelled differently [30, 31], such as with a Wright-Fisher model [32, 31], or with an approximation such as the logistic model used to describe myeloid leukemia [6]. As emphasised by some authors [50, 39], a large proportion of genetic alterations occur before adulthood. Therefore it seems possible that some routes to cancer could be determined prior to adulthood, with genetic mutations and epigenetic changes in childhood either favouring or inhibiting the possible paths by which adult cancers could arise. If this led to a given cancer type occurring with a small number of sufficiently different incident rates, then it might be observable in a population's incidence data as a mixture of distributions.

Changing micro-environment: Another potential limitation of the model described in Section 5 is that the time to failure at each step is independent of the other failure times, and of the time at which that step becomes at risk. If the tissue micro-environment is changing with time, then this assumption fails, and the failure rate at each step is dependent on the present time. This prevents the factorisation of the Laplace transform used in Eqs. 13-15, that led to Eq. 16 for failure via m sequential steps. We can explore the influence of a changing micro-environment with a perturbative approximation. The simplest example is to allow the $\{\mu_j\}$ in the Armitage-Doll

model to depend linearly on the time $\sum_{k=1}^{j} t_k$ at which step j is at risk. Then the Armitage-Doll approximation of $f_j(t_j) \simeq \mu_j$ for $\mu_j t_j \ll 1$, is replaced by

$$f_j(t_j|t_{j-1},\ldots,t_1) \simeq \mu_{j0} + \mu_{j1} \sum_{k=1}^{j} t_k \tag{57}$$

The calculation in Section 5 is modified, with,

$$P(T_1 = t_1,\ldots,T_m = t_m) = f_m(t_m|t_{m-1},\ldots,t_1)\ldots f_2(t_2|t_1)f_1(t_1) \tag{58}$$

giving,

$$\begin{aligned} P(T_1 = t_1,\ldots,T_m = t_m) &= \Pi_{j=1}^{m} \left(\mu_{j0} + \mu_{j1} \textstyle\sum_{k=1}^{j} t_k\right) \\ &= a_0 + \textstyle\sum_{j=1}^{m} a_j t_j^{m-j+1} \end{aligned} \tag{59}$$

with $a_0 = \Pi_{j=1}^{m} \mu_{j0}$, and $\{a_j\}$ being sums of products of $j-1$ factors from $\{\mu_{j0}\}$ and $m-j+1$ factors from $\{\mu_{k1}\}$. Replacing $\Pi_{j=1}^{m} f_j(t_j)$ in Eqs. 13 and 14, with the right-side of Eq. 59, and evaluating the m integrals then gives,

$$\mathcal{L}\,[f] = \frac{a_0}{s^m} + \sum_{j=1}^{m} a_j \frac{\Gamma(m-j+2)}{s^{m-j+2}} \frac{1}{s^{m-1}} \tag{60}$$

with solution,

$$f(t) = a_0 \frac{t^{m-1}}{\Gamma(m)} + \sum_{j=1}^{m} a_j \frac{\Gamma(m-j+2)}{\Gamma(2m-j+1)} t^{2m-j} \tag{61}$$

If the tissue micro-environment is changing rapidly enough that a term $a_j t_j^{2m-j}$ becomes comparable to or larger than $a_0 t^{m-1}$, then the solution to Eq. 61 can behave like a larger power of time than the usual $m-1$ for m rate-limiting steps. It is even possible for the incidence rate to slow or even *decrease*, if coefficients in Eq. 61 are negative. The example illustrates that if the micro-environment modifies cancer risk and is changing over a person's lifetime, then it has the potential to strongly influence the observed rate of cancer incidence. The argument can be repeated with less generality or greater sophistication, e.g. expanding the coefficients μ_i in the terms $\exp(-\mu_j t_j)$ that appear in Moolgavkar's model. Such models will have a complex relationship between their coefficients that might make them identifiable from cancer incidence data. This goes beyond the intended scope of this paper.

10 Summary

The purpose of this article is to provide a simple mathematical framework to describe existing multi-stage cancer models, that is easily adaptable to model events such as failure of complex systems, cascading disasters, and the onset of disease. The key formulae are Eqs. 1, 4, and 16 or equivalently 18, and a selection of analytical results are given to illustrate their use. Limitations of the multi-stage model are discussed in Sections 8 and 9. The examples in Section 6 can be combined in numerous ways to construct a wide range of models. Together the formulae are intended to provide a comprehensive toolkit for developing conceptual and quantitative models to describe failure, disaster, and disease.

Appendix

Derivation of Eq. 23 The solution of Eq. 16 can be written in terms of multiple definite integrals that are sometimes easier to evaluate or approximate than directly evaluating Eq. 16. It is equivalent to expressing the solution as multiple convolutions using Eq. 21, and changing variables appropriately. The equation is obtained by Taylor expanding all functions before taking their Laplace transform, inverting the Laplace transform of the product of all terms (which is easy to do for the powers of time that appear in a Taylor expansion), then using a product of Beta functions to factorise and re-sum the resulting expression. In mathematical notation,

$$\begin{aligned}\mathcal{L}^{-1}\left\{\Pi_{j=1}^{m}\mathcal{L}\left[f_j(t_j)\right]\right\} &= \mathcal{L}^{-1}\left\{\Pi_{j=1}^{m}\mathcal{L}\left[\textstyle\sum_{n_j=0}^{\infty} f_{j,n_j}\frac{t^{n_j}}{n_j!}\right]\right\}\\ &= \mathcal{L}^{-1}\left\{\textstyle\sum_{n_1=0}^{\infty}\cdots\sum_{n_m=0}^{\infty}\frac{f_{1,n_1}}{s^{n_1+1}}\cdots\frac{f_{m,n_m}}{s^{n_m+1}}\right\}\\ &= \textstyle\sum_{n_1=0}^{\infty}\cdots\sum_{n_m=0}^{\infty} f_{1,n_1}\cdots f_{m,n_m}\frac{t^{-1+\sum_{i=1}^{m}(n_i+1)}}{\Gamma(\sum_{i=1}^{m}(n_i+1))}\end{aligned} \tag{62}$$

where $f_{i,n_j} = \partial^{n_j} f_i(t_i)/\partial t^{n_j}|_{t_i=0}$. Now noting that the Beta function has,

$$\int_0^1 u^{m-1}(1-u)^{n-1}du = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \tag{63}$$

we can write,

$$\frac{1}{\Gamma\left(\sum_{i=1}^{m}(n_i+1)\right)} = \frac{1}{\Gamma\left(\sum_{i=1}^{m-1}(n_i+1)\right)}\frac{1}{\Gamma\left(n_m+1\right)}\int_0^1 y_{m-1}^{-1+\sum_{i=1}^{m-1}(n_i+1)}(1-y_{m-1})^{n_m} \tag{64}$$

Repeatedly using this gives,

$$\begin{aligned}\tfrac{1}{\Gamma\left(\sum_{i=1}^{m}(n_i+1)\right)} &= \left(\Pi_{i=1}^{m}\tfrac{1}{\Gamma(n_i+1)}\right)\textstyle\int_0^1 dy_1\cdots\int_0^1 dy_{m-1}y_1^0y_2^1y_3^2\ldots y_{m-1}^{m-2}\times\\ &\quad (1-y_1)^{n_2}\ldots(1-y_{m-1})^{n_m}\times(y_1\ldots y_{m-1})^{n_1}(y_2\ldots y_{m-1})^{n_2}\ldots y_{m-1}^{n_m-1}\end{aligned} \tag{65}$$

Using Eq. 65 to replace $1/\Gamma\left(\sum_{i=1}^{m}(n_i+1)\right)$ in Eq. 62, and grouping terms,

$$\begin{aligned}\mathcal{L}^{-1}\left\{\Pi_{j=1}^{m}\mathcal{L}\left[f_j(t_j)\right]\right\} &= t^{-1+m}\textstyle\int_0^1 dy_1\cdots\int_0^1 dy_{m-1}y_1^0y_2^1y_3^2\ldots y_{m-1}^{m-2}\times\\ &\quad\left(\textstyle\sum_{n_1=0}^{\infty} f_{1,n_1}\frac{t^{n_1}(y_1\ldots y_{m-1})^{n_1}}{\Gamma(n_1+1)}\right)\\ &\quad\left(\textstyle\sum_{n_1=0}^{\infty} f_{2,n_2}\frac{t^{n_2}(1-y_1)^{n_2}(y_2\ldots y_{m-1})^{n_2}}{\Gamma(n_2+1)}\right)\\ &\quad\ldots\\ &\quad\left(\textstyle\sum_{n_m=0}^{\infty} f_{m,n_m}\frac{t^{n_m}(1-y_{m-1})^{n_m}}{\Gamma(n_m+1)}\right)\end{aligned} \tag{66}$$

The m Taylor series can now be resummed to give,

$$\begin{aligned}\mathcal{L}^{-1}\left\{\Pi_{j=1}^{m}\mathcal{L}\left[f_j(t_j)\right]\right\} = t^{-1+m}\textstyle\int_0^1 dy_1\cdots\int_0^1 dy_{m-1}y_1^0y_2^1y_3^2\ldots y_{m-1}^{m-2}\times\\ f_1\left(ty_1\ldots y_{m-1}\right)f_2\left(t(1-y_1)(y_2\ldots y_{m-1})\right)f_3\left(t(1-y_2)(y_3\ldots y_{m-1})\right)\ldots f_m\left(t(1-y_{m-1})\right)\end{aligned} \tag{67}$$

For example, taking $m = 2$ gives,

$$f(t) = t\int_0^1 dy_1 f_1(ty_1)f_2(t(1-y_1)) \tag{68}$$

as we could have got from the convolution formula after a simple change of variables.

Eq. 67 might equivalently be regarded as a generalisation of a Schwinger/Feynmann parameterisation, with,

$$\begin{aligned}\Pi_{j=1}^{m}g_j(s) &= \textstyle\int_0^1 dy_1\cdots\int_0^1 dy_{m-1}y_1^0y_2^1y_3^2\ldots y_{m-1}^{m-2}\times\\ &\quad\mathcal{L}\left[t^{m-1}\mathcal{L}^{-1}\left\{g_1(s)\right\}\left(ty_1\ldots y_{m-1}\right)\right.\\ &\quad\mathcal{L}^{-1}\left\{g_2(s)\right\}\left(t(1-y_1)y_2\ldots y_{m-1}\right)\\ &\quad\ldots\\ &\quad\left.\mathcal{L}^{-1}\left\{g_m(s)\right\}\left(t(1-y_{m-1})\right)\right]\end{aligned} \tag{69}$$

For example, taking $g_j(s) = 1/(s+a_j)^{p_j}$ and noting that $\mathcal{L}^{-1}\left\{1/(s+a_j)^{p_j}\right\} = t^{p_j-1}e^{-a_j t}/\Gamma(p_j)$, then we get,

$$\begin{aligned}\Pi_{j=1}^{m}\tfrac{1}{(s+a_j)^{p_j}} \quad &= \tfrac{\Gamma\left(\sum_{i=1}^{m} p_i\right)}{\Pi_{i=1}^{m}\Gamma(p_i)} \int_0^1 dy_1 \cdots \int_0^1 dy_{m-1} y_1^0 y_2^1 \ldots y_{m-1}^{n_m-1} \\ &(y_1 \ldots y_{m-1})^{p_1-1}\left((1-y_1)y_2 \ldots y_{m-1}\right)^{p_2-1} \cdots \\ &(y_{m-1}(1-y_{m-2}))^{p_{m-1}-1}(1-y_{m-1})^{p_m-1} \\ &\frac{1}{[s+(a_1 y_1 \ldots y_{m-1}+a_2(1-y_1)y_2 \ldots y_{m-1}+\cdots+a_m(1-y_{m-1})]^{\sum_{i=1}^{m} p_i}}\end{aligned} \tag{70}$$

Taking $s = 0$, $m = 2$, and $p_j = 0$ for all j, gives the most well-known form, with [28],

$$\frac{1}{a_1 a_2} = \int_0^1 \frac{dy_1}{(a_2 y_1 + (1-y_1)a_1)^2} \tag{71}$$

The identity Eq. 70 can be confirmed by writing the denominator as $(A_m)^m$, with,

$$A_m = (a_m(1-y_{m-1}) + y_{m-1}A_{m-1}(y_1, \ldots, y_{m-2})) \text{ and } A_1 = a_1 \tag{72}$$

and integrating with respect to each of y_{m-1} to y_1 in turn. For example, using the substitution $u = y_{m-1}/(1+\alpha_{m-1}y_{m-1})$ with $\alpha_{m-1} = (A_{m-1} - a_m)/a_m$ and integrating between $u = 0$ and $u = 1/(1+\alpha_{m-1})$, the integrand becomes $(1/a_m)(1/{A_{m-1}}^m)$. Repeating this for y_{m-1} to y_1 confirms the identity.

Acknowledgments

Thanks to Benjamin Cairns and Andrii Rozhok for helpful comments. This research was funded by a fellowship from the Nuffield Department of Population Health, University of Oxford, and with support from Cancer Research UK (grant no. C570/A16491).

References

[1] C. O. Nordling, "A new theory on the cancer-inducing mechanism", *British Journal of Cancer*, vol. 7, no. 1, pp. 68-72, 1953.

[2] P. Armitage and R. Doll, "The age distribution of cancer and a multi-stage theory of carcinogenesis," *British Journal of Cancer*, vol. 8, no. 1, pp. 1–12, 1954.

[3] R. Peto, "Epidemiology, multistage models, and short-term mutagenicity tests," *International Journal of Epidemiology*, vol. 45, no. 3, pp. 621–637, 2016.

[4] S. H. Moolgavkar, "Commentary: Multistage carcinogenesis and epidemiological studies of cancer," *International Journal of Epidemiology*, vol. 45, no. 3, pp. 645–649, 2016.

[5] E. G. Luebeck and S. H. Moolgavkar, "Multistage carcinogenesis and the incidence of colorectal cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 23, pp. 15095–15100, 2002.

[6] F. Michor, Y. Iwasa, and M. A. Nowak, "The age incidence of chronic myeloid leukemia can be explained by a one-mutation model," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 40, pp. 14931–14934, 2006.

[7] R. Meza, J. Jeon, S. H. Moolgavkar, and E. G. Luebeck, "Age-specific incidence of cancer: Phases, transitions, and biological implications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 42, pp. 16284–16289, 2008.

[8] M. P. Little, "Cancer models, genomic instability and somatic cellular darwinian evolution," *Biology Direct*, vol. 5, p. 19, 2010.

[9] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. B. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[10] I. Martincorena, P. J. Campbell, "Somatic mutation in cancer and normal cells", *Science*, vol. 349, no. 6255, pp. 1483–1489, 2015.

[11] D. J. Ashley, "The two "hit" and multiple "hit" theories of carcinogenesis", *Br J Cancer* vol. 23, no. 2, pp. 313–328, 1969.

[12] A. G. Knudson, "Mutation and cancer: statistical study of retinoblastoma", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 68, no. 4, pp. 820–823, 1971.

[13] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[14] S. H. Moolgavkar, "Multistage theory of carcinogenesis and age distribution of cancer in man," *Journal of the National Cancer Institute*, vol. 61, no. 1, pp. 49–52, 1978.

[15] A. Ai-Chalabi, A. Calvo, A. Chio, S. Colville, C. M. Ellis, O. Hardiman, M. Heverin, R. S. Howard, M. H. B. Huisman, N. Keren, P. N. Leigh, L. Mazzini, G. Mora, R. W. Orrell, J. Rooney, K. M. Scott, W. J. Scotton, M. Seelen, C. E. Shaw, K. S. Sidle, R. Swingler, M. Tsuda, J. H. Veldink, A. E. Visser, L. H. van den Berg, N. Pearce, "Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study", *Lancet Neurology*, vol. 13, no. 11, pp. 1108–1113, 2014.

[16] A. Chio, L. Mazzini, S. D'Alfonso, L. Corrado, A. Canosa, C. Moglia, U. Manera, E. Bersano, M. Brunetti, M. Barberis, J. H. Veldink, L. H. van den Berg, N. Pearce, W. Sproviero, R. McLaughlin, A. Vajda, O. Hardiman, J. Rooney, G. Mora, A. Calvo, A. Al-Chalabi, "The multistep hypothesis of ALS revisited The role of genetic mutations", *Neurology*, vol. 91, no. 7, pp. E635–E642, 2018.

[17] P. Corcia, H. Blasco, S. Beltran, C. Andres, P. Vourc'h, P. Couratier, "In ALS, a mutation could be worth two steps", *Revue Neurologique*, vol. 174, no. 10, pp. 669–670, 2018.

[18] S. Licher, K. D. van der Willik, E. J. Vinke, P. Yilmaz, L. Fani, S. B. Schagen, M. A. Ikram, M. K. Ikram, "Alzheimer's disease as a multistage process: an analysis from a population-based cohort study" *Aging*, vol. 11, no. 4, pp. 1163–1176, 2019.

[19] A. AghaKouchak, L. S. Huning, O. Mazdiyasni, I. Mallakpour, F. Chiang, M. Sadegh, F. Vahedifard, and H. Moftakhari, "How do natural hazards cascade to cause disasters?," *Nature*, vol. 561, no. 7724, pp. 458–460, 2018.

[20] D. Collett, *Modelling Survival Data in Medical Research*, vol. Third Edition. CRC Press, 2015.

[21] C. Tomasetti, B. Vogelstein, "Variation in cancer risk among tissues can be explained by the number of stem cell divisions", *Science*, vol. 347, no. 6217, pp. 78–81, 2015.

[22] L. Nunney, "Size matters: height, cell number and a person's risk of cancer", *Proceedings of the Royal Society B-Biological Sciences*, vol. 285, no. 1889, art. no. 20181743, 2018.

[23] S. Wu, S. Powers, W. Zhu, and Y. A. Hannun, "Substantial contribution of extrinsic risk factors to cancer development," *Nature*, vol. 529, no. 7584, pp. 43–47, 2016.

[24] X. X. Zhang, H. Frohlich, D. Grigoriev, S. Vakulenko, J. Zimmermann, and A. G. Weber, "A simple 3-parameter model for cancer incidences," *Scientific Reports*, vol. 8, 2018.

[25] P. Calabrese and D. Shibata, "A simple algebraic cancer equation: calculating how cancers may arise with normal mutation rates," *BMC Cancer*, vol. 10, 2010.

[26] P. Armitage, R. Doll, and M. C. Pike, "Somatic mutation," *British Medical Journal*, vol. 1, no. 5436, pp. 723–, 1965.

[27] L. Soto-Ortiz and J. P. Brody, "A theory of the cancer age-specific incidence data based on extreme value distributions," *Aip Advances*, vol. 2, no. 1, 2012.

[28] S. Weinberg, *The Quantum Theory of Fields*, vol. 1. Cambridge University Press, 1995.

[29] M. Greaves, C. C. Maley, "Clonal evolution in cancer", *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.

[30] P. M. Altrock, L. L. Liu, F. Michor, "The mathematics of cancer: integrating quantitative models", *Nature Reviews Cancer*, vol. 15, no. 12, pp. 730–745, 2015.

[31] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, F. Markowetz, "Cancer Evolution: Mathematical Models and Computational Inference", *Systematic Biology*, vol. 64, no. 1, pp E1-E25, 2015.

[32] N. Beerenwinkel, T. Antal, D. Dingli, A. Traulsen, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, M. A. Nowak, "Genetic progression and the waiting time to cancer", *Plos Computational Biology*, vol. 3, no. 11, pp 2239–2246, 2007.

[33] J. Cairns, "Mutation selection and natural-history of cancer", *Nature*, vol. 225, no. 5505, pp. 197–200, 1975.

[34] P. C. Nowell, "The clonal evolution of tumor cell populations", *Science*, vol. 194, no. 4260, pp. 23–28, 1976.

[35] L. M. F. Merlo, J. W. Pepper, B. J. Reid, C. C. Maley, "Cancer as an evolutionary and ecological process", *Nature Reviews Cancer*, vol. 6, no. 12, pp. 924–935, 2006.

[36] R. A. Gatenby, "A change of strategy in the war on cancer", *Nature*, vol. 459, no. 7246, pp. 508–509, 2009.

[37] C. Tomasetti, L. Li, B. Vogelstein, "CANCER ETIOLOGY Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention", *Science*, vol. 355, no. 6331, pp. 1330–1334, 2017.

[38] M. J. Bissell, W. C. Hines, "Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression", *Nature Medicine*, vol. 17, no. 3, pp. 320-329, 2011.

[39] A. I. Rozhok, J. DeGregori, "Toward an evolutionary model of cancer: Considering the mechanisms that govern the fate of somatic mutations", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 29, pp. 8914–8921, 2015.

[40] C. C. Maley, A. Aktipis, T. A. Graham, A. Sottoriva, A. M. Boddy, M. Janiszewska, A. S. Silva, M. Gerlinger, Y. Yuan, K. J. Pienta, K. S. Anderson, R. Gatenby, C. Swanton, D. Posada, C. Wu, J. D. Schiffman, E. S. Hwang, K. Polyak, A. R. A. Anderson, J. S. Brown, S. Joel, M. Greaves, D. Shibata, "Classifying the evolutionary and ecological features of neoplasms", *Nature Reviews Cancer*, vol. 17, no. 10, pp. 605–619. 2017.

[41] R. A. Gatenby, "Is the Genetic Paradigm of Cancer Complete?", *Radiology*, vol. 284, no. 1, pp. 1–3, 2017.

[42] M. Lopez-Lazaro, "Cancer etiology: Variation in cancer risk among tissues is poorly explained by the number of gene mutations", *Genes Chromosomes & Cancer*, vol. 57, no. 6, pp. 281–293, 2018.

[43] M. Lopez-Lazaro, "The stem cell division theory of cancer", *Critical Reviews in Oncology Hematology*, vol. 123, pp. 95–113, 2018.

[44] R. A. Gatenby, A. S. Silva, R. J. Gillies, B. R. Frieden, "Adaptive Therapy", *Cancer Research*, vol. 69, no. 11, pp. 4894–4903, 2009.

[45] A. S. Silva, Y. Kam, Z. P. Khin, S. E. Minton, R. J. Gillies, R. A. Gatenby, "Evolutionary Approaches to Prolong Progression-Free Survival in Breast Cancer", *Cancer Research*, vol. 72, no. 24, pp. 6362–6370, 2012.

[46] P. M. Enriquez-Navas, Y. Kam, T. Das, S. Hassan, A. Silva, P. Foroutan, E. Ruiz, G. Martinez, S. Minton, S. R. J. Gillies, G. A. Gatenby, "Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer", *Science Translational Medicine*, vol. 8, no. 327, pp. 1–8, 2016.

[47] J. S. Zhang, J. J. Cunningham, J. S. Brown, R. A. Gatenby, "Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer", *Nature Communications*, vol. 8, no. 1816, 2017.

[48] R. J. Gillies, R. A. Gatenby, "Hypoxia and metabolism - Opinion - A microenvironmental model of carcinogenesis", *Nature Reviews Cancer*, vol. 8, no. 1, pp. 56–61, 2008.

[49] M. J. Bissell, W. C. Hines, "Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression", *Nature Medicine*, vol. 17, no. 3, pp. 320–329, 2011.

[50] S. Horvath, "DNA methylation age of human tissues and cell types", *Genome Biology*, vol. 14, no. 10, pp. R115, 2013.