THE UNIVERSITY
of EDINBURGH

Cancer recurrence times and early detection from branching process models

Stefano Avanzini

Doctor of Philosophy
University of Edinburgh
2019

Lay Summary

Detecting cancer early is recognized as one of the most effective strategies to improve prognosis and chances of survival. However, while screening and diagnostic technologies are advancing, their effectiveness relies on a quantitative understanding of the complex biological processes underlying the development of a tumor. Suitable theoretical frameworks to investigate and describe these processes are often provided by mathematical models, whose results can then be reinterpreted in the context of cancer evolution.

In this thesis we consider two mathematical descriptions for the growth of tumors and related cellular populations. The first one leads to estimates for the first time that a metastasis generated by a primary lesion becomes detectable. The second presents instead an abstract representation for an emerging and promising type of screening tests called liquid biopsies. Quantitative features of these models are also compared with relevant clinical data. Furthermore, both approaches are based on stochastic mathematical tools that are introduced at the beginning of this thesis.

Abstract

Cancer is among the leading causes of death worldwide. While primary tumors are often treated effectively, they can spawn secondary cancers called metastases which dramatically decrease chances of survival. In order to develop successful therapies, it is thus crucial to estimate the time until metastases appearance and improve our ability to detect primary tumors before metastases are generated. The estimation of the time to cancer recurrence depends on the dynamics of tumor growth and metastases seeding. For early detection, promising results have recently been obtained with liquid biopsies, id est the analysis of specific biomarker levels in blood samples. This thesis investigates these problems by studying mathematical models of cancer evolution and liquid biopsies based on the theory of branching processes.

Firstly, we consider first passage times to a given size in branching birth-death processes. We derive their probability distribution and first moments conditioned on non-extinction, comparing the results obtained for supercritical, critical and subcritical processes. Such results for hitting times are presented both in exact form and in their asymptotic limit for large sizes. In this limit we show that their probability distribution asymptotically converges to extreme value types.

Second, we present a semi-stochastic model of cancer recurrence. The primary tumor is described by a deterministically growing population of cells initiating metastases at a rate proportional to its size. Each metastasis is then modelled by a branching birth-death process with the same net growth rate. In this framework we discuss several features of the time to cancer relapse, defined as the first time that any metastasis reaches a given detectable size. We apply this model to different cancer types and compare its predictions with data collected from clinical literature.

Third, we present a multi-type branching process model of biomarker shedding. We focus on the case of circulating tumor DNA fragments shed in the bloodstream by both cancerous and healthy cells. We model the population of tumor cells as a supercritical branching birth-death process and take the healthy cells population to be constant in size. As DNA fragments cannot reproduce or divide, their amount is described by a pure death process with immigration. By applying this model, we provide quantitative estimates for the number of circulating tumor DNA fragments detectable in a blood sample, conditioned on the primary tumor size. Comparing our estimates with clinical observations we then discuss the potential of liquid biopsies for early cancer detection.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text. Where the work was done in collaboration with others, I have made a significant contribution. In particular, Chapter 4 is joint work with Johannes Reiter. This work has not been submitted for any other award or professional qualification.

21st August 2019 Stefano Avanzini

Acknowledgements

For the past four years, I would first of all like to thank my supervisor Tibor Antal. His guidance and constant encouragement have been fundamental in my learning experience as a PhD student. I am also grateful to my second supervisor, Michal Branicki, for his help and support. Special thanks go to Johannes Reiter for the opportunity he has given me to work together.

Further, I wish to acknowledge the School of Mathematics of the University of Edinburgh for providing the financial support for my PhD.

I would like to thank my fellow students, Michael Nicholson and David Cheek, for many research and non-research related discussions. My gratitude also goes to all the people that made the workplace a friendly and enjoyable environment, in particular Rory Mills-Williams, Diletta Martinelli, Martina Lanini, Roberto Fringuelli and Andrea Appel.

I would also like to thank the researchers met at conferences and workshops, and especially Annalisa Iuorio, for many helpful conversations.

I am extremely grateful to all the friends in Edinburgh that made these four years a very special time, in particular to Gabriele Piconi, Cathy Mitchell, Yair Fosado and all the LL Galaxy and Pleasance basketball communities.

Special thanks also go to all my friends and family in Italy, for their encouragement throughout this journey and for always being there for me, despite the distance.

Finally, a heartfelt thank you goes to Cecilia, for her love and support and for helping me making these years a time to remember.

Contents

List of Figures

List of Tables

Chapter 1

Background

1.1 Introduction

Early detection of cancer leads to higher chances of survival and, in general, better prognosis. However, modern screening techniques are not yet capable of reliably diagnosing a tumor before the appearance of symptoms or the formation of metastases [169]. Similarly, available estimates for the time to cancer recurrence still lack the required accuracy [189]. This thesis investigates these limitations by studying stochastic models of cancer evolution. In particular, the aim of this work is to develop mathematical tools that can help improve detection strategies for primary and secondary tumors. Our models are based on the theory of branching processes and birth-death processes, and are analyzed both qualitatively - by highlighting the mathematical features of the model - and quantitatively - by comparing model estimates with relevant clinical data.

Cancer evolution is an extremely complex biological process that is influenced by a multitude of factors. For example, the course of the disease varies widely across patients with different age, sex, lifestyle, genetic heritage or between tumors of different histological type, stage, aggressiveness. For this reason, cancer research focusses both on studying newly discovered aspects of tumor development and on improving our understanding of known cancer evolutionary dynamics. In line with these efforts, our work addresses topics related to cancer detection by presenting two models: the first is concerned with the long-studied problem of estimating the time to cancer recurrence, while the second discusses the potential for early detection of recently developed screening techniques based on liquid biopsies.

The first models pertaining the study of cancer considered the volume growth of a tumor over time and trace back to the 1920s [16]. Since then the number of mathematical applications to cancer research has increased enormously, providing insightful theoretical descriptions for several traits of tumor evolution, for example tumorigenesis [196], cancer growth [16], metastasis formation [172], drug resistance [114], intra-tumor heterogeneity [54] and many others. For historical notes and reviews on some of the most important and recent contributions in this field see e.g. [16, 11, 23].

In particular, in order to account for and characterize some of the variabilities intrinsic to cancer development, a growing number of stochastic models have been proposed, starting from the pioneering work of Armitage and Doll on carcinogenesis [17, 18]. In this context, a framework that has found many applications in cancer modelling is based on the so-called Luria-Delbrück experiment. In their Nobel prize winning work [127], Max Delbrück and Salvador Luria developed a model to study the emergence of subpopulations resistant to a toxin in a bacterial colony. They hypothesized that these subpopulations are initiated stochastically by the original bacteria through mutations at reproduction, and modelled the growth of both the original and mutant populations as deterministic exponential functions. Several generalizations of their model were later introduced to include stochastic growth of one or both populations (see [221, 37] for a review).

Our description of the time to cancer relapse is based on one of the semi-stochastic generalizations (that is on the assumption of a deterministic seeding intensity and a stochastic mutant growth) and it is presented in Chapter 3 of this thesis. More precisely, we model a primary tumor growth as a specified function of time, which stochastically initiates distant metastases. The evolution of each of these metastases is then represented by a branching birth-death process with positive net growth rate. Within this model we focus on the first time that any of the initiated metastases becomes detectable, which we define as the relapse of the disease. We also infer the model parameters for different cancer types, and analyze the corresponding predictions for several quantities of clinical interest. Our model can thus be employed to compute estimates for these quantities, as well as to gain insight into the dynamics of metastasis formation and growth. Both these quantitative and qualitative results can assist clinical practice, for example by informing on the expected time to relapse of the disease if a patient has no visible metastases at presentation, or on the effects of a delay in the time of primary tumor surgical resection.

In the model above, the derivation of the probability distribution for the time to cancer recurrence relies on results for hitting times to a certain size in branching birth-death processes. For this reason, in this thesis we first discuss the distributions of these times, which are the main subject of Chapter 2. Such distributions spawn from a broad mathematical literature that highlights non-trivial relationships between their exact form and special families of orthogonal polynomials and between their asymptotic limit for large sizes and extreme value theory.

An intuitive prediction of our model for the time to cancer relapse is that the earlier a primary tumor is detected and treated, the smaller are the chances of developing metastases. Driven by this observation, we then started to investigate screening methods for the early detection of cancer. Metastases form as cancer cells disseminate from the primary tumor and succeed to establish a secondary lesion in the surrounding tissue or elsewhere. In general, tumors are known to release different biological products in the circulating system. Some of these substances, or biomarkers, can thus be searched for in blood samples from patients and their analysis might in turn provide information on the tumor they were shed by. Similar blood tests, called liquid biopsies, offer several advantages in comparison to other screening techniques, but their potential and limitations - especially for early cancer detection - still need to be entirely evaluated.

In this context, we propose a fully stochastic model of biomarker shedding. We describe a population of cancerous cells as a supercritical branching birth-death process which releases biomarker molecules at a given rate. Overall, the biomarker is thus shed in the bloodstream at a stochastically varying rate. Each of the shed molecules is then modelled by a branching pure death process. Some biomarkers can be simultaneously released by a population of non-malignant cells, for example by a benign lesion. We include this factor in our model, and assume the size of the healthy population to stay constant over time. Overall, the biomarker is thus shed in the bloodstream at a stochastically varying rate from cancer cells and at a constant rate from healthy cells. In Chapter 4 of this thesis we apply this framework to circulating tumor DNA and quantify the dependence between tumor size and biomarker level. These results are then employed to assess the potential of liquid biopsies as screening tests in clinical practice.

As highlighted throughout this introduction, our models are based on similar settings where the evolution of biological populations is described by branching birth-death processes. Many biologically relevant questions arising in our anal-

ysis can be answered in terms of specific properties of these processes, and in particular by characterizing the time taken by the modelled population to reach a given number of cells. These standard and non-standard features represent the main tools for our investigation and will thus be discussed separately before introducing the two applied models.

With these motivations in mind, we now provide a short outline of this thesis.

1.1.1 Thesis outline

The rest of this thesis is organized as follows. In the remainder of this section we provide a basic introduction to the main mathematical tools employed in our work. Specifically, Section 1.2.1 presents classical notions related to branching birth-death processes - with and without immigration - and about the inversion of probability generating functions. In Section 1.2.2 we discuss fundamental concepts of extreme value theory, while Section 1.2.3 briefly introduces non-homogeneous Poisson processes and Cox processes.

In Chapter 2 we define hitting times to a given state in branching birth-death processes and derive their probability distributions. We discuss separately the cases of supercritical, critical and subcritical processes, and derive both exact distributions and their asymptotic limit for large sizes.

Chapter 3 is about our model of metastasis formation and growth. Based on a semi-stochastic generalization of the Luria-Delbrück model, we provide a formal definition of the time to cancer relapse. Applying the results previously described for branching birth-death processes, we then derive the probability distribution of such a time. We discuss many qualitative features of our model, especially in the case where the primary tumor is resected at a given time. Then, we compare our theoretical results with clinical data for five different cancer types.

In Chapter 4 we present our model of biomarker shedding. Here, the evolution of a primary tumor releasing a specific biomarker in the bloodstream is described as a two-type branching process. Our derivations, that are based on the computation of probability generating functions for these two processes, are then applied to the study of circulating tumor DNA and its use in liquid biopsies for the early detection of cancer.

In Chapter 5 we then summarize our results and discuss some open questions.

1.2 Mathematical preliminaries

This section provides a brief and informal introduction to mathematical concepts that are extensively used throughout this thesis. For each of these topics we then point out more detailed and exhaustive references. Additionally, this introduction is based on classical tools in probability theory and stochastic analysis, including generating functions, Markov chains and Kolmogorov equations. These areas are not covered here and are extensively discussed in many books such as [166, 103].

1.2.1 Branching birth-death processes

A birth-death process $(Z_t)_{t\geq 0}$ ([8]) is defined as a continuous time Markov chain on the non-negative integers, whose transition probabilities satisfy, as $\Delta t \to 0$,

$$\mathbb{P}(Z_{t+\Delta t} = j \mid Z_t = i) = \begin{cases} \alpha_i \Delta t + o(\Delta t) & j = i+1\,, \\ \beta_i \Delta t + o(\Delta t) & j = i-1\,, \\ 1 - (\alpha_i + \beta_i)\Delta t + o(\Delta t) & j = i\,, \\ o(\Delta t) & \text{otherwise}\ . \end{cases} \tag{1.1}$$

Birth-death processes have been extensively used in the last decades to describe biological and physical phenomena ([44, 144, 160]). In the former context, Z_t often represents the number of individuals in a population alive at time t. When Z_t is made of i individuals, α_i and β_i denote the rates at which births and deaths occur, respectively. The evolution of biological populations is frequently modelled under the additional assumption that individuals reproduce and die independently of each other with the same birth and death rates. In this case, the transition rates in equation (1.1) depend linearly on the population size, i.e. $\alpha_i = i\alpha$ and $\beta_i = i\beta$. Under the same hypotheses, we also notice that the progenies generated by each individual evolve as independent and identically distributed processes. By labelling the individuals alive at time t as $1, 2, \ldots, Z_t$, this feature can be formally written as

$$Z_{t+s} = \sum_{i=1}^{Z_t} Z_s^{(i)}\,, \tag{1.2}$$

where $Z_s^{(i)}$ denotes the size of the i-th progeny after time s. The property above characterizes a broad family of stochastic processes known as branching processes [20]. As we just showed this family includes birth-death Markov chains with rates proportional to the population size, which are therefore denominated branching birth-death processes.

For the branching birth-death process Z_t defined before, α and β are called birth and death rate, respectively. Denote $\lambda = \alpha - \beta$ the net growth rate of the process. When $\lambda > 0, \lambda = 0$ or $\lambda < 0$ the process is called supercritical, critical or subcritical, respectively. For any given initial condition $Z_0 = n_0$, we also denote $\mathcal{Z}^{(n_0)}(x,t) = \sum_{k=0}^{\infty} \mathbb{P}(Z_t = k \mid Z_0 = n_0)\, x^k$ the probability generating function of Z_t. Thanks to the independence of progenies in branching processes, for every $i \geq 1$ we have [20]

$$\mathcal{Z}^{(i)}(x,t) = \left[\mathcal{Z}^{(1)}(x,t)\right]^i . \tag{1.3}$$

An explicit expression of $\mathcal{Z}^{(1)}(x,t)$ can be derived from the backward Kolmogorov equations (see e.g. [8]) for Z_t and it is given by

$$\mathcal{Z}^{(1)}(x,t) = \begin{cases} \frac{(\alpha x-\beta)e^{-\lambda t}-\beta(x-1)}{(\alpha x-\beta)e^{-\lambda t}-\alpha(x-1)} & \text{if } \alpha \neq \beta\,, \\ \frac{1-(\alpha t-1)(x-1)}{1-\alpha t(x-1)} & \text{if } \alpha = \beta\,. \end{cases} \tag{1.4}$$

By combining equations (1.3) and (1.4) we immediately find an expression also for $\mathcal{Z}^{(n_0)}(x,t)$. From here, we compute the first moments of Z_t which yield

$$\mathbb{E}[Z_t \mid Z_0 = n_0] = \begin{cases} n_0 e^{\lambda t} & \text{if } \alpha \neq \beta\,, \\ n_0 & \text{if } \alpha = \beta \end{cases} \tag{1.5}$$

and

$$\mathrm{Var}(Z_t \mid Z_0 = n_0) = \begin{cases} n_0 \frac{\alpha+\beta}{\lambda} e^{\lambda t}\left(e^{\lambda t} - 1\right) & \text{if } \alpha \neq \beta\,, \\ 2n_0\alpha t & \text{if } \alpha = \beta\,. \end{cases} \tag{1.6}$$

One of the fundamental questions related to a branching or a birth-death process is if and when the process will go extinct. For a branching birth-death process Z_t, let us denote $\Omega_t = \{Z_t > 0\}$ and $\Omega_\infty = \{Z_t > 0 \text{ for all } t\}$ the events of Z_t survival up to time t and eventual survival (or non-extinction), respectively. The probability of the former event is equal to

$$\mathbb{P}(\Omega_t \mid Z_0 = n_0) = 1 - \mathcal{Z}^{(n_0)}(0,t) = \begin{cases} 1 - \left(\frac{\beta-\beta e^{-\lambda t}}{\alpha-\beta e^{-\lambda t}}\right)^{n_0} & \text{if } \alpha \neq \beta\,, \\ 1 - \left(\frac{\alpha t}{1+\alpha t}\right)^{n_0} & \text{if } \alpha = \beta\,. \end{cases} \tag{1.7}$$

As t goes to infinity, we then find

$$\mathbb{P}(\Omega_\infty \mid Z_0 = n_0) = \begin{cases} 1 - \left(\frac{\beta}{\alpha}\right)^{n_0} & \text{if } \alpha > \beta\,, \\ 0 & \text{if } \alpha \leq \beta\,. \end{cases} \tag{1.8}$$

Let us now focus on a branching birth-death process starting with one individual, and denote $q = \beta/\alpha$. The generating function $\mathcal{Z}^{(1)}(x,t)$ can be analytically inverted (see later paragraph) to find the probability mass function of Z_t. For $\alpha \neq \beta$ we find

$$\mathbb{P}(Z_t = k \mid Z_0 = 1) = \begin{cases} \frac{q}{\mathcal{S}(t)} & \text{if } k = 0\,, \\ \left(1 - \frac{q}{\mathcal{S}(t)}\right)(\mathcal{S}(t) - 1)\,\mathcal{S}(t)^{-k} & \text{if } k \geq 1\,, \end{cases} \tag{1.9}$$

where

$$\mathcal{S}(t) = \frac{1 - qe^{-\lambda t}}{1 - e^{-\lambda t}}\,.$$

By taking the limit as $\lambda \to 0$ we find instead the probability mass function for the critical case

$$\mathbb{P}(Z_t = k \mid Z_0 = 1) = \begin{cases} \frac{\alpha t}{1+\alpha t} & \text{if } k = 0\,, \\ \frac{1}{(1+\alpha t)^2}\left(\frac{\alpha t}{1+\alpha t}\right)^{k-1} & \text{if } k \geq 1\,. \end{cases} \tag{1.10}$$

When branching birth-death processes are employed to model the development of biological populations, it is often important to study the size of these populations at relatively large times. To investigate this problem we will exploit a classical property of branching processes [20]. Notice that the branching-birth-death process Z_t divided by its expected value, $\frac{Z_t}{\mathbb{E}[Z_t]} = Z_t e^{-\lambda t}$, is a non-negative martingale. Hence, by virtue of the martingale convergence theorem [61] there exists a non-negative random variable W such that

$$\lim_{t\to\infty} Z_t e^{-\lambda t} = W\,. \tag{1.11}$$

The distribution of W can be derived from the probability generating function of Z_t, and it is explicitly given by (see again [20])

$$\mathbb{P}(W \leq x) = q + (1-q)\left(1 - e^{-(1-q)x}\right), \quad x \geq 0\,. \tag{1.12}$$

From this expression we see that $W = 0$ if and only if Z_t goes extinct at some time t, which happens with probability q. Hence, we find that

$$\mathbb{P}(W \leq x \mid \Omega_\infty) = \mathbb{P}(W \leq x \mid W > 0) = 1 - e^{-(1-q)x}\,. \tag{1.13}$$

Branching birth-death processes with immigration

The evolution of biological populations can be greatly influenced by factors other than the births and deaths of their individuals, and in particular by dynamics of migration to and immigration from surrounding environments. To deal with similar situations, here we introduce branching birth-death processes with immigration [8]. These are defined as continuous time Markov chains Z_t on the non-negative integers whose transition probabilities satisfy, as $\Delta t \to 0$,

$$\mathbb{P}(Z_{t+\Delta t} = j \mid Z_t = i) = \begin{cases} (\nu + i\alpha)\Delta t + o(\Delta t) & j = i+1\,, \\ i\beta\Delta t + o(\Delta t) & j = i-1\,, \\ 1 - i(\alpha+\beta)\Delta t + o(\Delta t) & j = i\,, \\ o(\Delta t) & \text{otherwise}\ . \end{cases} \tag{1.14}$$

Let $\mathcal{Z}^{(n_0)}(x,t)$ denote the probability generating function for such a process, conditioned on $n_0 \geq 1$ initial individuals. Following steps similar to those described in the previous section, an explicit expression for $\mathcal{Z}^{(n_0)}(x,t)$ can be derived also in this case [8]

$$\mathcal{Z}^{(n_0)}(x,t) = \begin{cases} \dfrac{(\alpha-\beta)^{\nu/\alpha}\left[\beta(e^{\lambda t}-1) - (\beta e^{\lambda t} - \alpha)x\right]^{n_0}}{\left[\alpha e^{\lambda t} - \beta - \alpha(e^{\lambda t}-1)x\right]^{n_0+\nu/\alpha}} & \text{if } \alpha \neq \beta\,, \\ \dfrac{\left[x + \alpha t(1-x)\right]^{n_0}}{\left[1+\alpha t(1-x)\right]^{n_0-\nu/\alpha}} & \text{if } \alpha = \beta\,. \end{cases} \tag{1.15}$$

In turn, from these expressions we can again derive the probability mass function of Z_t and its first moments. In particular, we find

$$\mathbb{E}[Z_t \mid Z_0 = n_0] = \begin{cases} \frac{e^{\lambda t}(\lambda n_0 + \nu) - \nu}{\lambda} & \text{if } \alpha \neq \beta\,, \\ n_0 + \nu t & \text{if } \alpha = \beta\,. \end{cases}$$

Probability mass functions and first moments from generating functions

In the previous paragraphs we have derived the probability mass function for the process Z_t by inverting its generating function. As similar inversion techniques will frequently recur in the next chapters, we briefly discuss them here. Let $(\chi_t)_{t\geq 0}$ be a stochastic process and denote $\mathcal{G}(x,t) = \sum_{k=0}^{\infty} \mathbb{P}(\chi_t = k)x^k$ its probability generating function. The mass function of χ_t follows by expanding $\mathcal{G}(x,t)$ around

zero and is formally given by [61]

$$\mathbb{P}(\chi_t = k) = \frac{1}{k!} \left. \frac{\partial^k \mathcal{G}(x,t)}{\partial x^k} \right|_{x=0} .$$

The expected value and variance of χ_t are instead equal to

$$\mathbb{E}[\chi_t] = \partial_x \mathcal{G}(x,t)_{|x=1}, \quad \mathrm{Var}(\chi_t) = \partial_x^2 \mathcal{G}(x,t)_{|x=1} + \mathbb{E}[\chi_t] - \mathbb{E}[\chi_t]^2 .$$

In a few cases, the k-th derivative of the generating function is explicitly computable, thus allowing to find an analytic formula for the probability mass function of χ_t. The expression for $\mathcal{Z}^{(1)}(x,t)$ given by equation (1.4) is an example of such generating functions and from it we directly derived the probabilities $\mathbb{P}(Z_t = k \mid Z_0 = 1)$ - see equations (1.9) and (1.10). For most generating functions, however, this analytical inversion is not feasible and we have to compute the probability mass function numerically. Algorithms for this numerical procedure are based on techniques for series expansions of analytical functions, which are implemented as `Series` in both Mathematica (here see also `NSeries`) and Maple. For more details on the numerical inversion of probability generating functions we refer to [1, 41].

Stochastic simulations

To obtain the stochastic simulations of branching birth-death processes shown in this thesis we employed the Gillespie algorithm [74], which can be used to simulate more general Markov processes as well. In short, this algorithm relies on the property that for a Markov chain in a given state j, the time to the next jump is exponentially distributed and the mean of this distribution is the reciprocal of the total rate γ_j at which the process leaves state j. To build a realization of the process it is thus sufficient to iteratively update the total rate γ_j, compute the next interevent time by sampling from an $\mathrm{Exp}(1/\gamma_j)$ distribution and then determine the chain transition from the jump probabilities.

1.2.2 The extreme value theory

Many results presented in this thesis rely on the probability distribution of first passage times to a given size in branching birth-death processes, and in particular on their asymptotic form for large sizes. As we will discuss in detail in next chapter, these passage times are closely related to probability distributions aris-

ing in order statistics and extreme value theory. Hence, we now recall some basic notions in these fields. Classical references for these topics are [47, 6, 162].

Let $X_1, X_2, \ldots, X_n$ be a family of continuous and real-valued random variables defined on a common probability space. For every event ω in this space we can then arrange the values $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ in nondecreasing order. This allows us to introduce new random variables $X_{i,n}$, defined by the condition

$$X_{1,n}(\omega) < X_{2,n}(\omega) < \cdots < X_{n,n}(\omega) \quad \text{for every } \omega\,.$$

The random variable $X_{k,n}$ is called the k-th order statistic of the original sample. Suppose now that X_i are i.i.d. with common cumulative distribution and density functions F and f, respectively. Then, the density function of the k-th order statistic $X_{k,n}$, denoted by $f_{k,n}$, is given by

$$f_{k,n}(x) = \frac{n!}{(k-1)!(n-k)!}\,(F(x))^{k-1}\,(1-F(x))^{n-k}\,f(x)\,.$$

Let us here mention a special property of the order statistics of exponential distributions. If the random variables X_i are i.i.d. Exp(α), we can use the previous results to show that for every k [139]

$$X_{(k)} \stackrel{d}{=} \frac{1}{\alpha}\sum_{i=1}^{k}\frac{E_i}{n-i+1}\,,$$

where E_i denote i.i.d. standard exponential random variables. Such a distributional identity is known as the Renyi representation for the order statistics of exponential random variables [161].

Extreme value theory is concerned with the asymptotic behaviour of order statistics when the sample size gets large. More precisely, consider again a family of i.i.d. random variables $X_1, X_2, \ldots, X_n$ with common distribution function F and denote the upper end point of F as $\omega(F) = \sup\{x : F(x) < 1\}$. Then, we have

$$\mathbb{P}(X_{n,n} \le x) = \mathbb{P}(X_1 \le x, \ldots, X_n \le x) = F^n(x)$$

and so $X_{n,n}$ converges almost surely to $\omega(F)$. The aim of extreme value theory is then to find normalizing constants a_n, b_n and non-degenerate distribution

functions $G(x)$ such that

$$\mathbb{P}\left(\frac{X_{n,n}-b_n}{a_n}\leq x\right)\longrightarrow G(x)\quad \text{as } n\to\infty\,. \tag{1.16}$$

If such a_n, b_n and $G(x)$ exist, then the distribution function F is said to belong to the domain of attraction of G. Furthermore, two limit functions G and G' are said to be of the same type if there exist constants a and b such that $G'(x)=G(ax+b)$ for all x. The Fisher-Tippett-Gnedenko theorem states that there exist only three types of limit probability distributions for the maxima of i.i.d. random variables. These types are given by the following cumulative distribution functions

$$\begin{aligned}
&\Lambda(x)=e^{-e^{-x}}\,,\\
&\Phi_\gamma(x)=\begin{cases}0 & \text{if } x<0\\ e^{-x^{-\gamma}} & \text{if } x\geq 0\end{cases}\qquad \text{for some } \gamma>0\,,\\
&\Psi_\gamma(x)=\begin{cases}e^{-(-x)^{\gamma}} & \text{if } x<0\\ 0 & \text{if } x\geq 0\end{cases}\qquad \text{for some } \gamma>0
\end{aligned}$$

and are equivalently referred to as type I, II and III or as Gumbel, Fréchet and Weibull types, respectively. Many tests have been developed to assess whether a given distribution function belongs to the domain of attraction of one of these three types. For references on these methods see [121, 48].

To conclude this section, we provide a few more details about the Gumbel distribution. First, note that the exponential distribution belongs to its domain of attraction. To see this, it is sufficient to check that equation (1.16) is satisfied by $F(x)=1-e^{-\alpha x}, G(x)=e^{-e^x}, a_n=\frac{1}{\alpha}$ and $b_n=-\frac{\log(n)}{\alpha}$. A direct generalization of the Fisher-Tippett-Gnedenko theorem (see [47]) shows that there exist three types of possible nondegenerate distributions also for the limit of the k-th largest of i.i.d. random variables. Using the definitions in [5, 154], we characterize one of these types by saying that a continuous random variable X with support $(-\infty,\infty)$ follows the k-th generalized Gumbel distribution with parameters $a\in\mathbb{R}$ and $b>0$, $X\sim \mathrm{Gumb}_k(a,b)$, if and only if its cumulative distribution function is of the form

$$\mathbb{P}(X\leq t)=G_k(a,b;t):=e^{-e^{-\frac{t-a}{b}}}\sum_{j=0}^{k-1}\frac{e^{-j\frac{t-a}{b}}}{j!}\,. \tag{1.17}$$

If $X \sim \mathrm{Gumb}_k(a, b)$, then we have

$$\mathbb{E}[X] = a - b\,\psi_0(k)\ , \quad \mathrm{Var}(X) = b^2\,\psi_1(k)\,, \tag{1.18}$$

where ψ_0 and ψ_1 denote the digamma and trigamma functions, respectively [95].

Let us here provide distinct notations for the Gumbel limit of the largest and smallest extremes of a sample. The former coincides with the previous general case when $k = 1$, while the latter follows by a small adaptation. We say that a continuous random variable X with support $(-\infty, \infty)$ follows a Gumbel distribution for the maximum with parameters $a \in \mathbb{R}$ and $b > 0$, $X \sim \mathrm{Gumb}_{\max}(a, b)$, if and only if its cumulative distribution function is of the form [95]

$$\mathbb{P}(X \leq t) = G(a, b; t) = \exp\left(-e^{-\frac{t-a}{b}}\right)\ . \tag{1.19}$$

This distribution generally characterizes the maximum of independent random variables with exponential tail, and so it is a right skewed distribution. Similarly, we say that a continuous random variable X with support $(-\infty, \infty)$ follows a Gumbel distribution for the minimum with parameters $a \in \mathbb{R}$ and $b < 0$, and we denote it $X \sim \mathrm{Gumb}_{\min}(a, b)$, if and only if its cumulative distribution function is of the form

$$\mathbb{P}(X \leq t) = 1 - G(a, b; t) = 1 - \exp\left(-e^{-\frac{t-a}{b}}\right)\ . \tag{1.20}$$

This distribution describes the minimum of independent random variables with an exponential "nose", and its density is left skewed.

Here, let us anticipate that in the rest of this thesis, with a little abuse of notation, we will use the same symbol $\sim$ to denote the asymptotic behaviour of a function and the distribution of a random variable. For example, we will encounter both expressions like $\mathbb{P}(X \leq t) \sim G(a, b; t)$ and $X \sim \mathrm{Gumb}_{\min}(a, b)$, whose respective meanings will be unambiguously specified by the context.

To conclude this section let us note that, as $\psi_0(1) = \gamma_E$ and $\psi_1(1) = \pi^2/6$, if $X_1 \sim \mathrm{Gumb}_{\max}(a_1, b_1)$ and $X_2 \sim \mathrm{Gumb}_{\min}(a_2, b_2)$, for $i = 1, 2$ we have

$$\mathbb{E}[X_i] = a_i + b_i\,\gamma_E, \quad \mathrm{Var}(X_i) = \frac{b_i^2\,\pi^2}{6}\,, \tag{1.21}$$

where $\gamma_E \approx 0.5772$ denotes the Euler-Mascheroni constant.

1.2.3 Two-type processes and stochastic seeding

The models presented in this thesis are based on the assumption that cells from the primary tumor stochastically seed a second population. Intuitively, if the number of the cancerous cells alive at time t is A_t and the seeding rate per cell is ν, then at time t individuals of the second population are generated at rate νA_t. The form of this rate therefore depends on how we model the size A_t of the primary tumor. In particular, when A_t is described by a deterministic function or by a stochastic process, the second population is seeded as a non-homogeneous Poisson process or as a Cox process, respectively. Here we briefly review some of the main properties of these processes. More details can be found in classical references such as [165, 176, 45].

Non-homogeneous Poisson processes

A stochastic process $(K_t)_{t\geq 0}$ taking values on the non-negative integers, such that $K_0 = 0$ and $K_t \geq K_s$ almost surely for every $s \leq t$, is called a counting process. Let $\nu(t) : [0, \infty) \to [0, \infty)$ be an integrable function. If a counting process $(K_t)_{t\geq 0}$ is such that $K_{t_1} - K_{s_1}$ and $K_{t_2} - K_{s_2}$ are independent for any pair of disjoint intervals $[s_1, t_1]$ and $[s_2, t_2]$, and additionally satisfies

$$\mathbb{P}(K_{t+\Delta t} - K_t = n) = \begin{cases} 1 - \nu(t)\Delta t + o(\Delta t) & \text{if } n = 0\,, \\ \nu(t)\Delta t + o(\Delta t) & \text{if } n = 1\,, \\ o(\Delta t) & \text{if } n \geq 2 \end{cases}$$

for every t, then it is called a non-homogeneous Poisson process with rate (or intensity) $\nu(t)$. This process gets its name from the fundamental property that for any interval $[s, t]$, the random variable $K_{t+s} - K_t$ is Poisson distributed with parameter $\int_t^{t+s} \nu(u)du$. In particular, we thus have

$$\mathbb{E}[K_t] = \int_0^t \nu(u)du\,.$$

Clearly, when the rate function $\nu(t)$ is constant, K_t is a standard time homogeneous Poisson process.

Now, suppose that $p_i(t) : \mathbb{R} \to [0, 1]$, for $i = 1, \ldots, n$, are n continuous functions such that $\sum_{i=1}^n p_i(t) = 1$ for every t. Then, we can split the non-

homogeneous Poisson process K_t into

$$K_t = \sum_{i=1}^{n} K_t^{(i)}\,, \tag{1.22}$$

where $K_t^{(i)}$ denotes a non-homogeneous Poisson process with intensity $\nu(t)p_i(t)$. This property is called the splitting or thinning property of Poisson processes.

To conclude this section, let us mention that the stochastic simulations of non-homogeneous Poisson processes presented in this thesis are obtained through the Çinlar algorithm. Briefly, such algorithm is based on the result that a set of random variables $\{\sigma_i\}_{i\in\mathbb{N}}$ corresponds to the arrival times in a non-homogeneous Poisson process K_t with expectation function $\Lambda(t) := \int_0^t \nu(u)du$ if and only if the random variables $\{\Lambda(\sigma_i)\}_{i\in\mathbb{N}}$ correspond to the arrival times of a homogenous Poisson process with rate one. This result was first proved in [35], and allows to simulate K_t by iteratively sampling from a standard exponential distribution and then inverting the expectation function $\Lambda(t)$.

Cox processes

Cox processes generalize non-homogeneous Poisson processes by letting the intensity function to be stochastic. More precisely, let $(\nu_t)_{t\geq 0}$ be a random function taking values on the non-negative real line. Suppose that $(\nu_t)_{t\geq 0}$ is almost surely locally integrable, i.e. such that $\int_a^b \nu_t dt < \infty$ with probability 1 for every closed and bounded real interval $[a, b]$. A stochastic process $(K_t)_{t\geq 0}$ is then called a Cox process if conditional on $(\nu_t)_{t\geq 0}$ it is a non-homogenous Poisson process with rate $(\nu_t)_{t\geq 0}$. In particular, we have that for any interval $[s, t]$ such K_t satisfies

$$\mathbb{P}(K_t - K_s = n \mid \nu(u) : s \leq u \leq t) = \frac{\left(\int_s^t \nu(u)du\right)^n}{n!} e^{-\int_s^t \nu(u)du}\,.$$

Chapter 2

Hitting times for branching birth-death processes

2.1 Introduction

The time taken by a stochastic process to reach a specified state (or set of states) for the first time is usually referred to as first passage (or hitting) time to such state. First passage times have been extensively studied since the beginning of the 20th century and, partly due to the generality of their definition, found applications in a great number of diverse fields. For example, they can be used to plan the execution of a financial operation when a stock price reaches a certain threshold, or to model the triggering of chemical reactions when a molecular system assumes specific configurations. For the main properties of first passage times and a survey on some of their applications (e.g. to electrostatics, reaction phenomena and finance) we refer to [134, 157].

Around the same years, stochastic processes started to be widely employed in biology. In this context, first passage times have allowed us to study features of molecular transcription and translation, cellular mutation and disease, organismic evolution and many other processes [40]. In particular, in the field of population dynamics hitting times are often used to describe the time taken by a population to get extinct, or to grow to a certain size. Such a mathematical description applies to different kinds of biological species whose evolutionary dynamics - depending on the species characteristics and on modelling assumptions - are well represented by specific stochastic processes. For example, branching birth-death processes as those introduced in Section 1.2.1 are often employed for the modelling of cellular populations. An accessible review of this and other bi-

ological applications of birth-death processes is provided in [144].

For this type of processes, the main results concerning first passage times were first obtained by Samuel Karlin and James McGregor in the late 1950s [98, 99, 102]. In particular, by exploiting the family of orthogonal polynomials associated with a birth-death Markov chain, they proved that the probability distribution of hitting times is completely determined by the eigenvalues of the infinitesimal generator matrix underlying the process. While this property represents a powerful tool to study a broad range of birth-death chains, tractable expressions for the required eigenvalues are often impossible to obtain, and so simpler quantities like the expectation of hitting times [155] or asymptotic forms of their distribution are investigated.

In this thesis, we model the evolution of populations of cancerous cells as branching birth-death processes. For our applications we are especially interested in the first time that such populations take to reach a large detectable size M. Furthermore, as the modelled population might not grow to size M before getting extinct, we aim to obtain the probability distribution of that time conditioned on absorption in M. With this motivation in mind, in this chapter we review classical results for first passage times in a branching birth-death process, conditioned on its survival until reaching a certain size. While the majority of these results are known, they have often been derived independently of each other and using different mathematical approaches. We thus aim to discuss the probability distributions of hitting times in a coherent and accessible fashion, allowing to gain insight into the fundamental properties of these random variables. In this vein, our main contribution constists in providing a mathematical intuition for the asymptotic behaviour of first passage times to a large size.

The rest of this chapter is organized as follows. In Section 2.2 we first define hitting times for birth-death processes. We then discuss some simple properties of branching pure birth processes, that can be used as a reference example for the subsequent derivations. In Section 2.3 we state the Karlin-McGregor theorem, and we exploit it to derive the exact probability distribution of the hitting time to a given size M. In order to do so, we consider separately noncritical and critical processes. Then, in Section 2.4 we present the asymptotic distribution of hitting times to a large size, distinguishing again between noncritical and critical cases. It is well known that for noncritical processes these asymptotic results are closely related to extreme value distributions. At the end of this section we thus provide

a mathematical explanation for such a connection, and discuss the validity of this argument for the critical case. Finally, Section 2.5 summarizes the material presented.

The derivations of the simpler mathematical results discussed in this chapter are presented below their statement. More involved proofs are instead shown in Appendix A, while for some classical ones we will refer to the relevant literature.

2.2 Preliminaries

In this section we provide the main definitions and notation that will be used throughout the chapter. Then, we introduce some of the mathematical tools necessary to study hitting times distributions by briefly discussing the special case of a branching pure birth process.

2.2.1 Definitions and notation

Let Z_t be a branching birth-death processes on the non-negative integers, as defined in Section 1.2.1. By definition these processes are skip-free, i.e. they do not allow for one-step transitions from a state to a non-neighbouring one. As a consequence, the first passage time of Z_t to a size M is formally defined as

$$T_M = \inf\{t > 0 : Z_t = M\}\,. \tag{2.1}$$

Depending on the realization of the process, this time can be finite or infinite. In particular, if Z_t gets extinct before it reaches size M we set $T_M = \infty$. For our applications we will often be interested in studying the probability distribution of T_M conditional on Z_t reaching size M in a finite time. For this reason, on top of the survival events Ω_t and Ω_∞ introduced in Section 1.2.1, we define

$$\Omega^j = \{Z_t = j \text{ for some } t \geq 0\}$$

as the event of Z_t survival up to size j, for every $j \geq 1$.

The probability distribution of the hitting time T_M also depends on the initial size of the process Z_t. In the following, we will thus express the conditioning on a given initial state (or population size) through the notation

$$\mathbb{P}_i(\,\cdot\,) = \mathbb{P}(\,\cdot \mid Z_0 = i)\,.$$

Moreover, our analysis of hitting times will occasionally require us to consider their distributions additionally conditioned on Z_t always staying above or below its initial size. To this purpose, we introduce the following notations

$$\mathbb{P}_{\underline{i}}(\cdot) = \mathbb{P}(\cdot \mid Z_0 = i,\ Z_s \geq i \text{ for all } s)\,,$$
$$\mathbb{P}_{\overline{i}}(\cdot) = \mathbb{P}(\cdot \mid Z_0 = i,\ Z_s \leq i \text{ for all } s)\,.$$

By recalling the definition of Z_t eventual survival we see in particular that

$$\mathbb{P}_{\underline{1}}(T_j \leq t) = \mathbb{P}(T_j \leq t \mid Z_0 = 1, \Omega_\infty)\,.$$

2.2.2 A reference example: branching pure-birth processes

Let Z_t be a branching pure birth process, i.e. a branching birth-death process with death rate $\beta = 0$. By definition, Z_t is a Markov chain representing the size of a population where each individual splits independently at rate α and cannot die. Hence, the lifetime of each individual in such a population is exponentially distributed with rate α. We define the inter-event times for Z_t as the random variables $Y_i := T_{i+1} - T_i$, where T_i denotes the first passage time to i (see equation (2.1)). Clearly, this definition implies

$$T_M = \sum_{i=n_0}^{M-1} Y_i\,. \tag{2.2}$$

Now, suppose that $Z_0 = 1$ and for every individual let us label its two children as 0 and 1, respectively. Then, every individual alive at a given time is uniquely associated to a binary sequence starting with a zero, which labels the first ancestor. Denote l_a the lifespan of the individual labelled by the binary sequence a. Because of the previous observations, we have that $l_a \sim \text{Exp}(\alpha)$ for every label a. Moreover, since we are assuming that $Z_0 = 1$, we have $T_1 = 0$ and hence

$$Y_1 = T_2 = l_{\{0\}} \sim \text{Exp}(\alpha)\,.$$

The second inter-event time is then determined by the minimum of the lifespans $l_{\{0,0\}}$ and $l_{\{0,1\}}$, which are again independent $\text{Exp}(\alpha)$ random variables. Therefore we have

$$Y_2 \sim \text{Exp}(2\alpha)\,.$$

Now, let us loosely refer to the length of a binary sequence a as the generation of the individual labelled by a. Without loss of generality, suppose that among the

two individuals in the second generation the one labelled by $\{0,1\}$ was the first one to divide, determining the time Y_2. Then, the lifetime left for the other individual in the second generation, $l_{\{0,0\}} - Y_2$, is still an $\mathrm{Exp}(\alpha)$ random variable thanks to the memoryless property of exponential distributions. As a consequence, we find

$$Y_3 = \min\left(l_{\{0,1,0\}}, l_{\{0,1,1\}}, l_{\{0,0\}} - Y_2\right) \sim \mathrm{Exp}(3\alpha)\,.$$

This argument is visualized in Figure 2.1, and can be extended to prove (e.g. by induction) that

$$Y_i \sim \mathrm{Exp}(i\alpha) \quad \text{for every } i \geq 1\,. \tag{2.3}$$

By joining equations (2.2) and (2.3) we thus see that for a branching pure birth process starting with $n_0 \geq 1$ individuals

$$\mathbb{P}_{\underline{n_0}}(T_M \leq t) = \mathbb{P}\left(\sum_{i=n_0}^{M-1} \mathrm{Exp}(i\alpha) \leq t\right)\,. \tag{2.4}$$

On the other hand, the parameters of the exponential random variables in equation (2.4) can be scaled out as follows

$$\sum_{i=n_0}^{M-1} \mathrm{Exp}(i\alpha) \overset{d}{=} \frac{1}{\alpha}\sum_{i=n_0}^{M-1} \frac{E_i}{i}\,,$$

where E_i are i.i.d. standard exponentials (i.e. $\mathrm{Exp}(1)$). The right hand side of the last equation is then recognized as the Rényi representation for the $(M-n_0+1)$-th order statistic from $M-1$ exponential random variables with parameter α (see Section 1.2.2). In other words, we have that for a branching pure birth process with birth rate α and n_0 initial individuals the first passage time T_M is distributed like the n_0-th largest out of $M-1$ $\mathrm{Exp}(\alpha)$ random variables. Here, let us recall that the order statistics coming from exponentially distributed samples belong to the (generalized) Gumbel domain of attraction (see again Section 1.2.2).

But what is the intuition behind T_M behaving like an order statistic in the first place? To answer this question, let us consider a branching pure death process Z_t^* with death rate α and such that $Z_0^* = M-1$. We denote T_i^* and Y_i^* the hitting time to size i and the i-th inter-event time for Z_t^*, respectively. A reasoning similar to the one we made for Z_t yields

$$T_0^* = \sum_{i=1}^{M-1} Y_i^* \sim \mathrm{Exp}(i\alpha)\,.$$

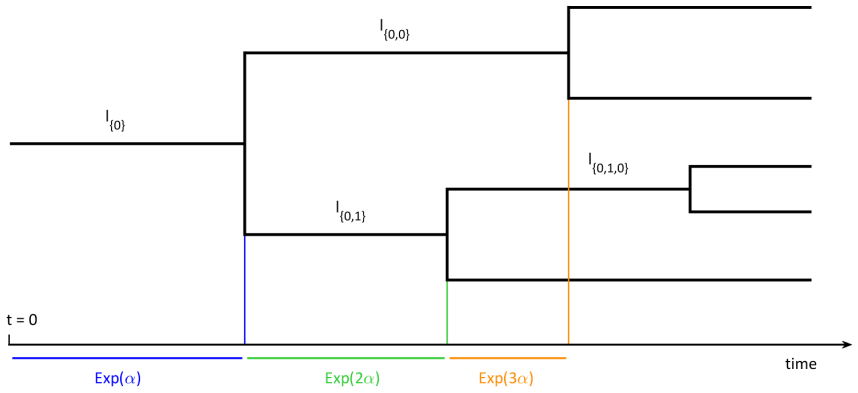

Figure 2.1: ***Inter-event times distributions in a pure birth process.***

Therefore, by applying without ambiguity the same notations introduced in Section 2.2.1 to Z_t^*, we find $\mathbb{P}_{\underline{1}}(T_M \leq t) = \mathbb{P}_{M-1}(T_0^* \leq t)$. But since T_0^* is the time to extinction of a pure death process, it coincides with the maximum lifespan of its initial individuals. Similarly T_1^*, that has the same distribution of the hitting time T_M conditional on $Z_0 = 2$, is equal to the second longest lifespan of Z_t^* initial individuals, and so on.

In the rest of this chapter we will exploit and generalize some of the arguments presented here to study the distribution of T_M for branching birth-death processes. To conclude this section let us mention that in the branching pure birth case, the asymptotic Gumbel limit of hitting times distributions was first noticed by Blackwell and Kendall [26] and then studied in great details by Waugh in a series of papers (see [201, 202, 204]). All these authors derived such a result by expressing T_M in terms of the so-called stochastic lag $S = -\frac{\log(W)}{\alpha}$, where W is the classic martingale limit for branching processes (see equation (1.11)).

2.3 Exact hitting times distributions

In this section we derive the exact probability distribution of the hitting time to size M for a branching birth-death process, conditioned on the process survival up to M. Such a distribution follows from a classical theorem that applies to a broader family of birth-death processes and that was originally shown by Samuel Karlin and James McGregor.

2.3.1 Karlin-McGregor theorem

Consider a birth-death process Z_t defined on the non-negative integers and with reflecting boundary at 0, i.e. such that

$$\mathbb{P}(\hat{Z}_1 = 1 \mid \hat{Z}_0 = 0) = 1,$$

where $\hat{Z}_n$ denotes the jump process associated to Z_t. The following theorem characterizes the probability distribution of the first passage time to M for such a process.

Theorem 2.3.1. *Let Z_t be a birth-death process on $\{0, \dots, M\}$ such that $Z_0 = 0$. If 0 is a reflecting boundary for the process and M is treated as an absorbing state, then the first passage time T_M is distributed as the sum of M exponential random variables whose parameters are the negatives of the non-zero eigenvalues of the underlying infinitesimal generator.*

Notice that to construct the infinitesimal generator the theorem refers to it is sufficient to truncate the generator of the birth-death chain Z_t between 0 and M and additionally set $\mathbb{P}(\hat{Z}_1 = M \mid \hat{Z}_0 = M) = 1$. The result above has been derived in different ways. As we mentioned before the first authors to state it and prove it were Karlin and McGregor [102]. Their proof relies on the theory of orthogonal polynomials associated to a birth-death process, developed by the same authors, that we will briefly introduce in the next paragraph. The first probabilistic proof of Theorem 2.3.1 was presented instead by Persi Diaconis and Laurent Miclo [51]. All these demonstrations are quite involved and we thus do not report them here. For more details we refer to the two papers cited above and in particular to the historical notes discussed in [51], that remark additional contributions. Notice, however, that from a modelling standpoint no clear interpretation of the exponential random variables in the theorem statement has been provided yet.

Before we discuss the application of Theorem 2.3.1 to branching birth-death processes, let us show how we can exploit it to compute explicitly the probability density function of T_M when the birth-death process Z_t starts with a given number $n_0 \geq 1$ of individuals.

Proposition 2.3.2. *Let Z_t be a birth-death process on $\{0, 1, 2, \dots\}$ such that 0 is a reflecting state. For every k, let $\boldsymbol{Q}_k$ be the infinitesimal generator of the process Z_t restricted between 0 and k, when $Z_0 = 0$ and k is treated as an absorbing*

state. Moreover, suppose that $0 < n_0 < M$ *and define* λ_i, $i = 1, \dots, M$ *and* μ_j, $j = 1, \dots, n_0$ *as the negatives of the non-zero eigenvalues of* $\boldsymbol{Q}_M$ *and* $\boldsymbol{Q}_{n_0}$, *respectively. Then, by denoting*

$$f_{n_0,M}(t) = \frac{d}{dt}\,\mathbb{P}(T_M \leq t \mid Z_0 = n_0)\,,$$

we have

$$f_{n_0,M}(t) = \sum_{k=1}^{M} \xi_k\, e^{-\lambda_k t}\,,$$

where

$$\xi_k = \frac{\lambda_1 \cdots \lambda_M}{\mu_1 \cdots \mu_{n_0}} \cdot \frac{\prod_{j=1}^{n_0}(\mu_j - \lambda_k)}{\prod_{\substack{i=1 \\ i \neq k}}^{M}(\lambda_i - \lambda_k)}\,. \tag{2.5}$$

The proof of this proposition is presented in Appendix A.1. We now provide a short introduction to the theory of orthogonal polynomials associated with a birth-death process.

Associated orthogonal polynomials

The relationship between birth-death processes and orthogonal polynomials was first highlighted in a series of seminal papers from Karlin and McGregor [98, 99, 100, 101, 102]. Their work proved that the transition probabilities of any given birth-death process can be expressed in terms of a family of polynomials and a Borel measure associated to the process. Since then, many other properties of a birth-death process have been described in terms of these polynomials, including the probability distributions of hitting times. In this paragraph we briefly review some of these topics, while for more details and some of the classical literature on the subject we refer to [104, 181, 132, 91, 194].

Let Z_t be a birth-death process on the non-negative integers as defined by equation (1.1). Its transition probabilities admit the following representation

$$\mathbb{P}(Z_t = j \mid Z_0 = i) = \pi_j \int_0^\infty e^{-xt} Q_i(x) Q_j(x) \psi(dx)\,,$$

where the coefficients π_j are given by

$$\pi_0 = 0, \quad \pi_n = \frac{\alpha_0 \alpha_1 \cdots \alpha_{n-1}}{\beta_1 \beta_2 \cdots \beta_n},$$

ψ is a Borel measure on $[0, \infty)$ of total mass 1 and $\{Q_n\}$ is a family of polynomials satisfying the recurrence relation

$$\begin{gathered} \alpha_n Q_{n+1}(x) = (\alpha_n + \beta_n - x) Q_n(x) - \beta_n Q_{n-1}(x), \quad n > 1, \\ \alpha_0 Q_1(x) = \alpha_0 + \beta_0 - x, \quad Q_0(x) = 1 \end{gathered} \tag{2.6}$$

and orthogonal with respect to ψ, i.e. such that

$$\int_0^\infty Q_i(x) Q_j(x) \psi(dx) > 0 \quad \text{if and only if} \quad i \neq j.$$

The measure ψ and the sequence $\{Q_n\}$ are called the spectral measure and the orthogonal polynomials associated with Z_t, respectively. Under additional assumptions it is possible to prove that the measure ψ is uniquely determined by the transition rates of the birth-death process. Several recurrence relations similar to that in equation (2.6) can be derived by applying simple algebraic transformations. As an example, consider the family of polynomials defined by

$$R_n(x) = (-1)^n \alpha_0 \cdots \alpha_{n-1} Q_n(-x) \tag{2.7}$$

for every $n \geq 0$. These polynomials satisfy (see e.g. [193])

$$\begin{gathered} R_{n+1}(x) = (-x - \alpha_n - \beta_n) R_n(x) - \alpha_{n-1} \beta_n R_{n-1}(x), \quad n \geq 1, \\ R_1(x) = -x - \alpha_0 - \beta_0, \quad R_0(x) = 1. \end{gathered} \tag{2.8}$$

Now, let us denote $\boldsymbol{Q}_n$ the infinitesimal generator of a birth-death process Z_t stopped at n. It is then easy to check that the recurrence relation (2.8) defines precisely the family of characteristic polynomials $R_n(x) = \det(\boldsymbol{Q}_n - x I_n)$, where I_n denotes the $n \times n$ identity matrix. The zeros of these polynomials thus coincide with the eigenvalues of $\boldsymbol{Q}_n$, which in turn provides the link between Theorem 2.3.1 and the family $\{Q_n\}$. Unfortunately, this family reduces to a known class of polynomials only in few cases. As we will see in the next section, one of these special cases is represented by critical branching birth-death processes and hence some of its properties will directly follow by the form of the associated orthogonal polynomials. The same properties for noncritical processes will instead require alternative derivations.

Conditional branching birth-death processes

Now, suppose that Z_t is a branching birth-death process and let us denote $\hat{Z}_t$ its embedded Markov chain (or associated jump process). Theorem 2.3.1 does not apply directly to this case, since 0 is an absorbing state. However, consider Z_t conditioned on survival up to size M, and denote $\tilde{p}_{i,j}$ the corresponding conditional one-step transition probabilities. In particular, we have

$$\tilde{p}_{1,0} = \mathbb{P}(\hat{Z}_1 = 0 \mid \hat{Z}_0 = 1, \Omega^M) = 0\,.$$

The conditional process thus reduces to a Markov chain defined on $\{1, \dots, M\}$ such that the lower end of its state space, i.e. 1, is a reflecting boundary. Theorem 2.3.1 can now be applied to such a process by simply relabelling the states through the map $i \mapsto i - 1$. The first passage time to M is therefore distributed as the sum of $M - 1$ exponential random variables, whose parameters depend on the conditional transition probabilities $\tilde{p}_{i,j}$. In the following sections we derive explicit expressions for these probabilities and discuss consequent properties of hitting times distributions.

2.3.2 Noncritical case

Let Z_t be a supercritical branching birth-death process with birth rate α and death rate β. If $Z_0 = 1$, its extinction probability is equal to $q = \beta/\alpha < 1$ (see equation (1.8)). Consider such a process restricted to $\{0, \dots, M\}$, for a fixed $M > 1$. Its conditional transition probabilities $\tilde{p}_{i,j} = \mathbb{P}(\hat{Z}_1 = j \mid \hat{Z}_0 = i, \Omega^M)$ are given by

$$\tilde{p}_{i,j} = \begin{cases} \dfrac{1 - q^{i+1}}{(1+q)(1-q^i)} & \text{if } j = i + 1\,, \\ \dfrac{q(1 - q^{i-1})}{(1+q)(1-q^i)} & \text{if } j = i - 1\,, \\ 0 & \text{otherwise }, \end{cases} \tag{2.9}$$

where $0 \leq i, j \leq M$. A full derivation of these expressions can be found for example in [203]. Notice that by conditioning on Ω^M we leave unchanged the rate of inter-event times, which for every individual of Z_t is equal to $\alpha + \beta = \alpha(1+q)$. Hence, if $\tilde{\boldsymbol{Q}}_M$ denotes the infinitesimal generator matrix of Z_t conditioned on Ω^M and $\tilde{q}_{i,j}$ its entries, the previous observation implies that $\tilde{q}_{i,i} = -i\alpha(1+q)$ for

every $0 \leq i \leq M$. When $i \neq j$ we have instead $\tilde{q}_{i,j} = \tilde{p}_{i,j} \cdot \tilde{q}_{i,i}$, and so we find

$$\tilde{q}_{i,j} = \begin{cases} i\alpha \, \dfrac{1-q^{i+1}}{1-q^i} & \text{if } j = i+1\,, \\ -i\alpha\,(1+q) & \text{if } j = i\,, \\ i\alpha\, q \, \dfrac{1-q^{i-1}}{1-q^i} & \text{if } j = i-1\,, \\ 0 & \text{otherwise }, \end{cases} \tag{2.10}$$

for $0 \leq i, j \leq M$. Furthermore, we observe that the expressions for $\tilde{q}_{i,j}$ and $\tilde{p}_{i,j}$ only depend on the states i and j and not on the final size M. Intuitively, this is due to the fact that each of these rates and probabilities govern the process simply conditional on reaching state $\max\{i,j\}$ before getting extinct. Hence, we have

$$\mathbb{P}(\hat{Z}_1 = j \mid \hat{Z}_0 = i, \Omega^M) = \mathbb{P}(\hat{Z}_1 = j \mid \hat{Z}_0 = i, \Omega^{\max\{i,j\}}) = \tilde{p}_{i,j}\,,$$

which also explains why we do not include M in their notation. As a consequence, we have that for every $n > 1$ the infinitesimal generator $\tilde{\boldsymbol{Q}}_n$ can be constructed from $\tilde{\boldsymbol{Q}}_{n-1}$ by simply adding to it the $(n+1)$-th row and column.

Now let us denote $R_n(x) = \det(\tilde{\boldsymbol{Q}}_n - xI_{n+1})$, where I_n represents the $n \times n$ identity matrix. The previous observation allows us to apply to these characteristic polynomials the same argument described in the previous section. Specifically, by expanding $R_n(x)$ over the two non-zero elements of the $(n+1)$-th row, $\tilde{q}_{n-1,n}$ and $\tilde{q}_{n,n}$, we can express it in terms of $R_{n-1}(x)$ and $R_{n-2}(x)$. The sequence $\{R_n(x)\}_n$ is then uniquely identified by the following recursive relation

$$\begin{gathered} R_0(x) = 1\,, \quad R_1(x) = -\alpha(1+q) - x\,, \\ R_n(x) = [-n\alpha(1+q) - x]\, R_{n-1}(x) - n(n-1)\alpha^2 q\, R_{n-2}(x) \end{gathered} \tag{2.11}$$

for all $n \geq 2$. Referring now to the discussion presented at the end of Section 2.3.1, we recall that the distribution of the hitting time T_M conditioned on Ω^M is determined - as a result of Theorem 2.3.1 - by the eigenvalues of the infinitesimal generator $\tilde{\boldsymbol{Q}}_{M-1}$. This property, together with the previous derivation, implies that

$$\mathbb{P}\left(T_M \leq t \mid Z_0 = 1, \Omega^M\right) = \mathbb{P}\left(\sum_{i=1}^{M-1} \mathrm{Exp}(-r_{M-1,i}) \leq t\right)\,,$$

where $r_{M-1,i}$ denote the $M-1$ zeros of the polynomial $R_{M-1}(x)$. These distributions, for different values of the rates α and β, are plotted in Figure 2.2. As an

example, we notice that when $Z_0 = 1$, the first passage time to 2 conditional on Ω^2 is distributed like the first inter-event time, as the process cannot transition to 0. Such a time is exponentially distributed with parameter $\alpha(1+q)$, which is equal to $-r_{1,1}$ (see equation (2.11)).

In general, however, many properties of the zeros $r_{M,i}$ are not known since the polynomials $R_n(x)$ identified by the recurrence relation (2.11) - or equivalently the associated orthogonal polynomials $Q_n(x)$ obtained by inverting equation (2.7) - do not correspond to classical families. Hence, while the discussion above yields the exact distribution of T_M, it is hard to derive its asymptotic limit for large M or other properties of the process Z_t from similar arguments.

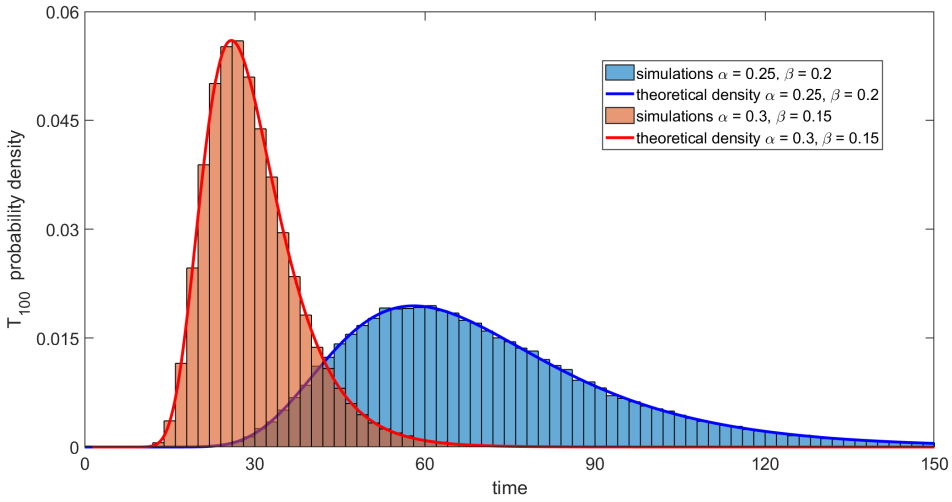

Figure 2.2: ***Conditional first passage time densities in a supercritical branching birth-death process.*** *The histograms represent the distribution of the hitting time to $M = 100$ obtained from 10^5 simulations of a supercritical branching birth-death process with the birth and death rates indicated in the legend and conditioned on reaching M. The solid lines show instead the corresponding conditional theoretical densities computed from Theorem 2.3.1. Already for $M = 100$ and different values of the birth and death rates, these densities perfectly match the simulated data.*

A different scenario occurs when studying critical branching birth-death processes. Before we address this latter case, let us conclude this section by mentioning a formula for the expected value of the hitting time T_M and how the previous results apply to subcritical branching birth-death processes.

Proposition 2.3.3. *Let Z_t be a supercritical branching birth-death process with birth rate α, death rate β and such that $Z_0 = n_0 \geq 1$. Then*

$$\mathbb{E}\left[T_M \mid Z_0 = n_0, \Omega^M\right] = \frac{1}{\alpha} \sum_{i=n_0}^{M-1} \frac{q^i}{(1-q^i)(1-q^{i+1})} \sum_{j=1}^{i} \frac{(1-q^j)^2}{jq^j}. \tag{2.12}$$

The proof of this result as well as a formula for the variance is provided in Appendix A.2.1.

Subcritical case

Let Z_t be a subcritical branching birth-death process with birth rate β and death rate α, where $\alpha > \beta$. Then, its transition probabilities conditional on survival up to M have the same form as in equation (2.9) (see e.g. [203]). The previous reasoning can thus be repeated and applied also to this case. In particular we find that, by conditioning on survival up to M, the hitting times T_M in two noncritical branching birth-death processes with birth and death rates swapped have the same probability distribution. A similar "reversal" argument will be explored more in depth in Section 2.4 to gain insight into the asymptotic behaviour for large sizes of these distributions.

2.3.3 Critical case

By adapting the argument employed in the noncritical case, we can derive the distribution of hitting times for critical branching birth-death processes as well. In this case, the family of orthogonal polynomials associated with these processes are well known and so the distribution provided by Theorem 2.3.1 is more tractable. In particular, we find the following

Proposition 2.3.4. *Let Z_t be a critical branching birth-death process with the same birth and death rates α and such that $Z_0 = 1$. Then, for every $M > 1$, the first passage time to M conditioned on survival up to M is distributed like a sum of independent exponential random variables with parameters $\alpha\lambda_{M-1,i}$, where $\lambda_{M-1,i}$ are the $M-1$ zeros of the associated Laguerre polynomial $L^{(1)}_{M-1}(\lambda)$.*

Sketch of the proof. Let us employ the same notations used in the previous section. We consider the critical branching birth-death process Z_t defined in the statement and restricted to $\{0, \ldots, M\}$. The one-step transition probabilities and the transition rates for such a process follow by taking the limit as $q \to 1$ in

equations (2.9) and (2.10), respectively. Hence we find

$$\tilde{p}_{i,j} = \begin{cases} \dfrac{i+1}{2i} & \text{if } j = i+1\,, \\ \dfrac{i-1}{2i} & \text{if } j = i-1\,, \\ 0 & \text{otherwise }, \end{cases} \qquad \tilde{q}_{i,j} = \begin{cases} \alpha(i+1) & \text{if } j = i+1\,, \\ -2\alpha i & \text{if } j = i\,, \\ \alpha(i-1) & \text{if } j = i-1\,, \\ 0 & \text{otherwise }, \end{cases}$$

where $0 \leq i, j \leq M$. The recurrence relation for the characteristic polynomials $R_n(x) = \det(\tilde{\boldsymbol{Q}}_n - xI_{n+1})$ is obtained by plugging the rates above in equation (2.8), or by taking again the limit as $q \to 1$ in equation (2.11), and is given by

$$R_n(x) = (-2\alpha n - x)R_{n-1}(x) - \alpha^2 n(n-1)R_{n-2}(x), \quad n \geq 2\,,$$
$$R_1(x) = -x - 2\alpha, \quad R_0(x) = 1\,.$$

Let us now set $\alpha = 1$. From equation (2.7), we see that in this case the orthogonal polynomials associated to Z_t are equal to $Q_n(x) = \frac{(-1)^n}{n!}R_n(-x)$, and thus satisfy the recurrence relation

$$nQ_n(x) = (2n - x)Q_{n-1}(x) - nQ_{n-2}(x), \quad n \geq 2\,,$$
$$Q_1(x) = 2 - x, \quad Q_0(x) = 1\,.$$

This last relation determines the associated Laguerre polynomials of index $\nu = 1$ (see for example [19]), defined as

$$L_M^{(1)}(x) := \sum_{j=0}^{M} (-1)^j \binom{M+1}{M-j} \frac{x^j}{j!}. \tag{2.13}$$

By construction, the zeros of $R_M(x)$ coincide with those of $L_M^{(1)}(x)$. Extending this argument to the case $\alpha \neq 1$, we find that $R_M(x) = (-1)^M \alpha^M M! L_M^{(1)}(-x/\alpha)$. In particular, this means that if $\lambda_{M,i}$ denotes the i-th zero of $L_M^{(1)}(x)$, the eigenvalues of $\tilde{\boldsymbol{Q}}_M$ are equal to $\alpha\lambda_{M,i}$. The rest of the proof then follows as in the noncritical case. □

The distribution of T_M for $M = 100$ and different values of α is visualized in Figure 2.3. As for the previous section, we conclude our analysis of the critical case by providing expressions for the conditional expectation of T_M.

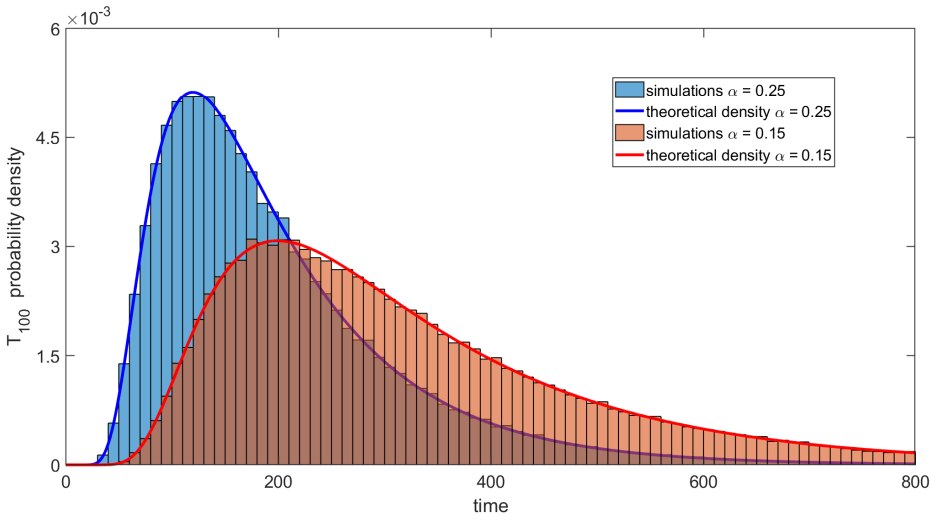

Figure 2.3: ***Conditional first passage time densities in a critical branching birth-death process.*** *The histograms show the distribution of the hitting time to* $M = 100$ *obtained from* 10^5 *simulations of a critical branching birth-death process with the birth rate indicated in the legend and conditional on absorption in* M*. The solid lines plot instead the corresponding conditional theoretical densities computed from Proposition 2.3.4. Already for* $M = 100$ *and different values of the birth rate, these densities perfectly match the simulated data.*

Proposition 2.3.5. *Let* Z_t *be a critical branching birth-death process with birth and death rates* α*. Then*

$$\mathbb{E}\left[T_M \mid Z_0 = n_0, \Omega^M\right] = \frac{M - n_0}{2\alpha}\,.$$

The derivation of this expression, as well as one for the variance, is presented in Appendix A.2.2. Such a proof follows the same argument used for the supercritical case. However, since in this case the parameters $\lambda_{M,i}$ are zeros of a known family of polynomials, the same result can be computed by summing up the reciprocals of $\alpha\lambda_{M-1,i}$ (and applying Proposition 2.3.2 if $n_0 > 1$). The value of this sum is then provided by Viete's formulas (see [195]).

2.4 Asymptotic hitting times distributions

In this section we discuss the asymptotic behaviour of the hitting time distributions derived before, and consider again separately noncritical and critical processes.

2.4.1 Noncritical case

Theorem 2.4.1. *Let Z_t be a supercritical branching birth-death process with birth rate α, death rate β and such that $Z_0 = 1$. Then, as M gets large, the distribution of the first passage time T_M asymptotically converges to*

$$\mathbb{P}_{\underline{1}}(T_M \leq t) \sim e^{-(1-q)Me^{-\lambda t}}, \tag{2.14}$$

where $\lambda = \alpha - \beta$ denotes the net growth rate of the process and $q = \beta/\alpha$ its extinction probability.

Proof. When $Z_0 = 1$, we know that the process $Z_t/e^{\lambda t}$ converges almost surely to a martingale limit (see equation (1.11)) as $t \to \infty$. Since $\lim_{M\to\infty} T_M \overset{a.s.}{=} \infty$, the same result implies that

$$\lim_{M\to\infty} \frac{M}{e^{\lambda T_M}} \overset{a.s.}{=} W\,. \tag{2.15}$$

The cumulative distribution of W is also known. In particular, we saw that for $Z_0 = 1$ and conditioned on Ω_∞, the random variable W follows an exponential distribution with parameter $1 - q$ (see equation (1.13)). Therefore, by joining this result and equation (2.15) we find

$$\begin{aligned}\mathbb{P}_{\underline{1}}(T_M \leq t) &= \mathbb{P}_{\underline{1}}\left(Me^{-\lambda T_M} \geq Me^{-\lambda t}\right) \\ &\sim 1 - \mathbb{P}_{\underline{1}}(W \leq Me^{-\lambda t}) = e^{-(1-q)Me^{-\lambda t}}\,.\end{aligned}$$ □

Figure 2.4 shows a comparison between the exact and asymptotic distributions of T_M, which become undistinguishable already for relatively small values of M. A similar approach to that used in the previous proof can also be employed to derive the asymptotic distribution of T_M conditional on $n_0 > 1$ initial individuals. Indeed, we find

Proposition 2.4.2. *Let Z_t be a supercritical branching birth-death process with birth rate α and death rate β. Then, for M large, the distribution of T_M conditioned on non-extinction and n_0 initial individuals asymptotically converges to*

$$\mathbb{P}_{n_0}(T_M \leq t \mid \Omega_\infty) \sim \sum_{j=0}^{n_0-1} c_j\, G_{n_0-j}\left(\frac{1}{\lambda}\log[(1-q)M], \frac{1}{\lambda}; t\right), \tag{2.16}$$

where $G_k(a, b; t)$ denotes the k-th generalized Gumbel distribution and

$$c_j := \frac{\binom{n_0}{j} q^j (1-q)^{n_0-j}}{1 - q^{n_0}}, \quad j = 0, \ldots, n_0 - 1\,.$$

This last proposition, whose detailed proof can be found in Appendix A.3.1, explicitly shows that when $n_0 > 1$ the conditional distribution of T_M asymptotically converges to a mixture of generalized Gumbel distributions with weights c_j. The cumulative distribution function of these distributions is given by equation (1.17).

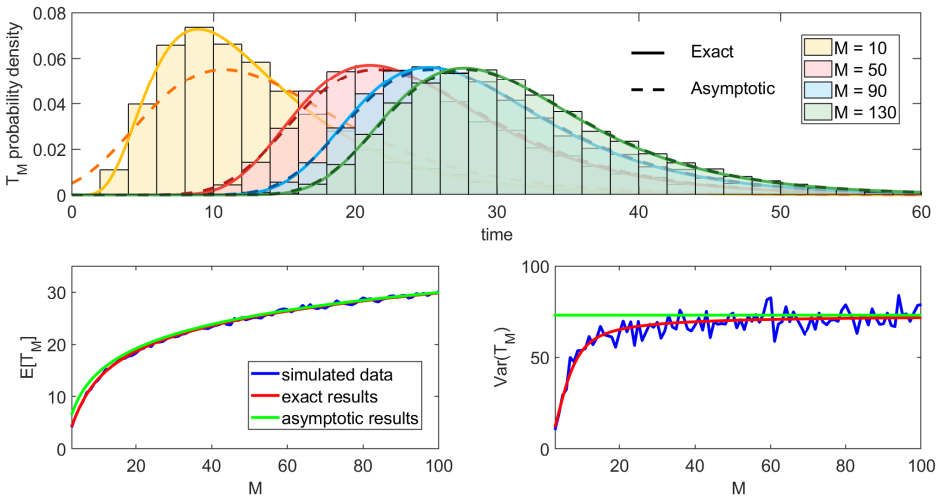

Figure 2.4: ***Comparison between exact and asymptotic distributions of hitting times in a supercritical branching birth-death process.*** *The histograms in the top panel show the distribution of the first passage time to* $M = 10, 50, 90, 130$ *obtained from* 10^5 *simulations of a supercritical branching birth-death process with birth rate* $\alpha = 0.3$ *and death rate* $\beta = 0.15$ *day*$^{-1}$*. In the same panel, solid lines represent the corresponding exact theoretical densities computed from Theorem 2.3.1, while the dashed lines plot the asymptotic Gumbel densities derived from equation* (2.14)*. The exact and asymptotic distributions differ for* M *very small and become almost undistinguishable around* $M = 100$*. This result is confirmed by the two bottom panels, which compare the simulated expectation and variance of* T_M *with their exact and asymptotic expression (see equations* (2.12) *and* (A.7)*, and Proposition 2.4.3 ,respectively).*

An expression equivalent to equation (2.16) was first obtained in [208] through a different mathematical derivation. As we pointed out at the end of Section 2.2.2, Waugh did instead exploit a martingale argument and applied it to study branching pure birth processes, but he did not carry out an explicit computation of $\mathbb{P}_{\underline{1}}(T_M \leq t)$ in the birth-death case. Finally, we notice that for $n_0 = 1$, the proof we have just showed has been recently employed in [53].

Following our presentation of the exact results for hitting times distributions in noncritical branching birth-death processes, we now conclude this section by providing the asymptotic mean and variance of these times for large sizes, and

with a brief discussion of the subcritical case.

Proposition 2.4.3. *Let Z_t be a supercritical branching birth-death process with birth rate α and death rate β. Then, asymptotically for M large, we have*

$$\begin{aligned}\mathbb{E}\left[T_M \mid Z_0 = n_0, \Omega^M\right] &\sim \frac{1}{\lambda}\log(M) + C_1\,,\\ \mathrm{Var}\left(T_M \mid Z_0 = n_0, \Omega^M\right) &\sim C_2\,,\end{aligned}$$

where C_1 and C_2 are two constants (with respect to M) given by

$$\begin{aligned}C_1 &= \frac{\log(1-q) - \sum_{j=0}^{n_0-1} c_j \psi_0(n_0-j)}{\lambda},\\ C_2 &= \frac{1}{\lambda^2}\left[\sum_{j=0}^{n_0-1} c_j\left[\psi_1(n_0-j) + (\psi_0(n_0-j))^2\right] - \left[\sum_{j=0}^{n_0-1} c_j\psi_0(n_0-j)\right]^2\right].\end{aligned}$$

In these expressions, the coefficients c_j are defined as in Proposition 2.4.2, while ψ_0 and ψ_1 denote the digamma and trigamma functions, respectively.

The proof of this result is shown in Appendix A.2.1. Notice that since $\psi_0(1) = \gamma_E$ and $\psi_1(1) = \pi^2/6$, Proposition 2.4.3 implies in particular that

$$\begin{aligned}\mathbb{E}\left[T_M \mid Z_0 = 1, \Omega^M\right] &\sim \frac{1}{\lambda}\log(M) + \frac{\log(1-q)+\gamma_E}{\lambda}\,,\\ \mathrm{Var}_1\left(T_M \mid Z_0 = 1, \Omega^M\right) &\sim \frac{\pi^2}{6\lambda^2}\,,\end{aligned}$$

where $\gamma_E \approx 0.5772$ is the Euler-Mascheroni constant.

Subcritical case

Recalling the results presented at the end of Section 2.3.2, we know that the distributions of hitting times for subcritical branching birth-death processes follows straightforwardly from the corresponding results for a supercritical process with rates swapped. The same argument can be used to derive the asymptotic behaviour of these distributions. In particular, we find that if Z_t is a subcritical branching birth-death process with birth rate β, death rate α and starting with one individual, then asymptotically for M large

$$\mathbb{P}_{\underline{1}}(T_M \leq t) \sim e^{-(1-q)Me^{-\lambda t}}\,.$$

In a similar fashion we can then exploit the previous results to obtain the distribution of T_M conditioned on $n_0 > 1$ initial individuals and its first moments.

2.4.2 Critical case

As shown by Proposition 2.3.4, the distribution of the first hitting time to M in a critical branching birth-death process depends entirely on the zeros of $L^{(1)}_{M-1}(\lambda)$. As M gets large, these zero converge to those of a Bessel function. Hence, by combining their asymptotic behaviour and Proposition 2.3.4 we find the following

Theorem 2.4.4. *Let Z_t be a critical branching birth-death process with birth and death rates α and starting with one individual. Then, as M becomes large, the first passage time T_M conditional on survival up to M is asymptotically distributed as $\sum_{i=1}^{M-1} \overline{X}_i$, where $\overline{X}_i$ are independent random variables such that*

$$\overline{X}_i \sim \text{Exp}\left(\frac{\alpha j_i^2}{4M}\right), \quad i = 1, \dots, M-1\,,$$

and j_i denotes the i-th positive zero of the Bessel function of the first kind $J_1(z)$.

The proof of Theorem 2.4.4 is presented in Appendix A.3.2. A comparison between this last result and the exact distribution of T_M is visualized by Figure 2.5, which highlights that the asymptotic approximation provided by Theorem 2.4.4 is extremely good for almost all values of M. Now, notice that since the random variables $\overline{X}_i$ in the statement of Theorem 2.4.4 are exponentially distributed, we can scale out their parameters as we did to derive the Rényi representation of hitting times in the branching pure birth case (see Section 2.2.2). In particular we have

$$\sum_{i=1}^{M-1} \overline{X}_i = \frac{4M}{\alpha} \sum_{i=1}^{M-1} X_i\,,$$

where $X_i \sim \text{Exp}(j_i^2)$ for every i. Hence, by joining this observation and Theorem 2.4.4 we find that, conditioned on $Z_0 = 1$ and Ω_∞,

$$\frac{\alpha T_M}{4M} \xrightarrow[M\to\infty]{d} \sum_{i=1}^{\infty} X_i\,.$$

As noticed by Pakes [148], the distribution provided in Theorem 2.4.4 was first investigated by Ciesielski and Taylor [42] in the context of diffusion limits for branching processes.

We also stress that in the critical case the hitting time T_M, conditioned on

Ω^M and $Z_0 = 1$, is distributed like a sum of exponentials both for finite M and in the asymptotic limit for M large. Hence, the asymptotic distribution of T_M conditioned on absorption in M and $n_0 > 1$ initial individuals follows by adapting Proposition 2.3.2 to the parameters of the exponential random variables in Theorem 2.4.4. The same reasoning applies to the noncritical case only when M is finite, and not in the asymptotic regime.

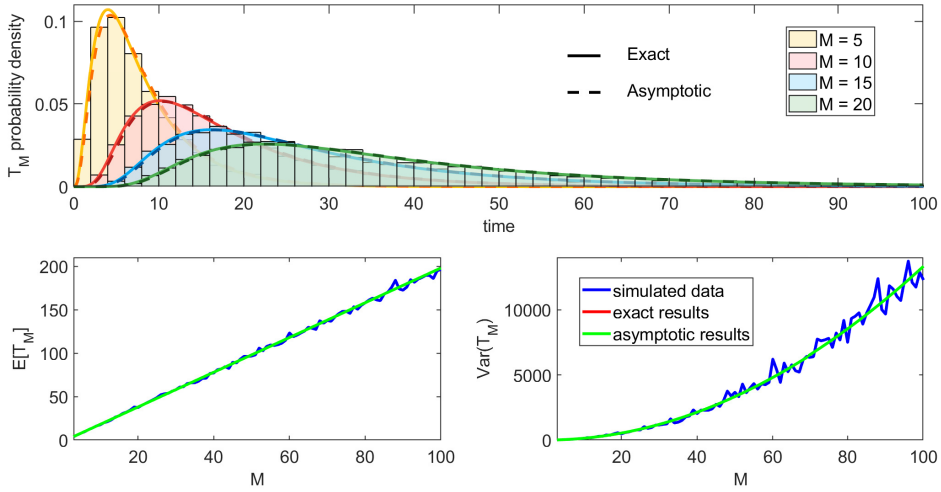

Figure 2.5: ***Comparison between exact and asymptotic distributions of hitting times in a critical branching birth-death process.*** *The histograms in the top panel show the distribution of the first passage time to $M = 5, 10, 15, 20$ obtained from 10^5 simulations of a critical branching birth-death process with birth rate $\alpha = 0.15$ day^{-1}. In the same panel, solid lines represent the corresponding exact theoretical densities computed from Proposition 2.3.4, while the dashed lines plot their asymptotic limit derived from Theorem 2.4.4. The exact and asymptotic distributions are very similar already for $M = 5$, and from $M \geq 10$ they cannot be distinguished. The goodness of the approximation provided by the asymptotic results in the critical case is also visualized in the two bottom panel, which show that the exact and asymptotic expressions for the mean and variance of T_M (see Proposition 2.3.5 and equation* (A.8)*, and equation* (2.17) *,respectively) perfectly match each other.*

Finally, the expression for the asymptotic mean of T_M can be derived from its exact forms given by Proposition 2.3.5. Looking also at the formula for the exact variance provided in the appendix (see equation (A.8)), we find

$$\mathbb{E}[T_M \mid \Omega^M] \sim \frac{M}{2\alpha}, \qquad \mathrm{Var}(T_M \mid \Omega^M) \sim \frac{M^2}{12\alpha^2}\,. \tag{2.17}$$

The same results can be obtained again through the parameters of the exponential random variables in Theorem 2.4.4. For this approach, sums of terms j_i^{-2p} for $p = 1, 2$ are discussed for example in [175].

2.4.3 Extreme value interpretation

In the previous section we showed that hitting times in noncritical branching birth-death processes asymptotically follow an extreme value distribution. However, our derivations do not clarify why these random variables behave like an extreme. In order to interpret such results, we will exploit an argument based on reversed birth-death processes, that we therefore introduce next.

Let Z_t be a branching birth-death process with birth rate α and death rate β. The reversed (sometimes also called dual) process Z_t^* is defined as a branching birth-death process with rates swapped, i.e. with birth rate β and death rate α. In the following we will add a superscript asterisk to the notations introduced so far every time we need to refer to quantities related to the reversed process. In particular, looking at Section 2.2.1, we denote

$$T_j^* = \inf\{t > 0 : Z_t^* = j\}$$

and

$$\begin{aligned}
\mathbb{P}_i^*(\cdot) &= \mathbb{P}(\cdot \mid Z_0^* = i)\,, \\
\mathbb{P}_{\underline{i}}^*(\cdot) &= \mathbb{P}(\cdot \mid Z_0^* = i,\ Z_s^* \geq i \text{ for all } s)\,, \\
\mathbb{P}_{\bar{i}}^*(\cdot) &= \mathbb{P}(\cdot \mid Z_0^* = i,\ Z_s^* \leq i \text{ for all } s)\,.
\end{aligned}$$

Noncritical case

Assume that Z_t is a supercritical branching birth-death process with birth rate α and death rate β. Let Z_t^* be the reversed process - which is therefore subcritical - and suppose that $Z_0^* = M$. The asymptotic distribution of the time to extinction of Z_t^* as M gets large has been extensively studied (see for example [147, 92, 67]). In our notation, it is given by

$$\mathbb{P}_M^*\left(T_0^* \leq t\right) \sim e^{-\left(\frac{\alpha-\beta}{\alpha}\right)Me^{-|\lambda^*|t}} = e^{-(1-q)Me^{-\lambda t}}\,, \tag{2.18}$$

where $q = \beta/\alpha$ and $\lambda = \alpha - \beta$ are the extinction probability and the net growth rate of the original process Z_t, respectively. We immediately notice that the expression in equation (2.18) coincides with the asymptotic distribution of T_M conditional on Ω^M and $Z_0 = 1$, as given by equation (2.14). On the other hand, we observe that by the branching property (see equation (1.2)), each of the initial M individuals of Z_t^* generate independent processes. Hence, in order for the Z_t^*

to get extinct, each of these M lineages need to die out. The time T_0^* is thus the maximum of the extinction times of independent and identically distributed processes, which explains why it follows an extreme value distribution.

Motivated by these observations, we now investigate the identity between equations (2.14) and (2.18). To this purpose, a key property is provided by the following.

Lemma 2.4.5. *For any birth-death process Z_t and any i, j,*

$$\mathbb{P}_{\underline{i+1}}(T_{j+1} \leq t) = \mathbb{P}_{\bar{j}}(T_i \leq t)\,.$$

This result was first shown by Sumita [180] and it implies in particular that if we consider a birth-death process starting with one individual and conditioned on non extinction, its first passage time to M has the same distribution of the time that an identical process starting with $M-1$ individuals and conditioned on no visits to M takes to get extinct. This argument can in turn be applied to prove the next proposition.

Proposition 2.4.6. *Let Z_t be a supercritical branching birth-death process with birth rate α and death rate β, and Z_t^* be the reversed process. Then*

$$\mathbb{P}_{\underline{1}}(T_M \leq t) = \mathbb{P}^*_{\overline{M-1}}(T_0^* \leq t)\,.$$

Proof. Following the discussion at the end of Section 2.3.2, we find that

$$\mathbb{P}_{\underline{1}}(T_M \leq t) = \mathbb{P}^*_{\underline{1}}(T_M^* \leq t)\,.$$

The statement then follows by directly applying Lemma 2.4.5. □

In light of this result, we observe that the identity between equations (2.14) and (2.18) is shown if we can prove that the asymptotic limit of $\mathbb{P}^*_{\overline{M-1}}(T_0^* \leq t)$ as M gets large coincides with that of $\mathbb{P}^*_M(T_0^* \leq t)$. To this purpose, we relate these latter two distributions by decomposing the time to extinction for Z_t^* starting from M individuals as

$$\mathbb{P}^*_M(T_0^* \leq t) = \mathbb{P}^*_M(R_M^* + X_0^* \leq t)\,. \tag{2.19}$$

Here $R_M^* = \sup\{t > 0 : Z_t^* = M\}$ is the last exit time from M and X_0^* denotes the extinction time "left". From the definition of R_M^* it follows that the only possible

transition after R_M^* is from M to $M-1$ and subsequently it is not possible to go back to M. Therefore, the distribution of X_0^* is exactly given by $\mathbb{P}^*_{\overline{M-1}}(T_0^* \leq t)$.

Lemma 2.4.7. *Let R_M^* be the last exit time from M defined above. Then, its distribution conditional on M initial individuals goes to zero in probability as M tends to ∞.*

The proof of this Lemma can be found in Appendix A.4. Finally, by joining the previous results we can prove the following.

Theorem 2.4.8. *Let Z_t be a branching birth-death process and let Z_t^* be its reversed process. Then, as M becomes large we have*

$$\mathbb{P}_{\underline{1}}(T_M \leq t) \sim \mathbb{P}^*_M(T_0^* \leq t)\,.$$

Proof. Lemma 2.4.7, together with equation (2.19), implies that

$$\mathbb{P}^*_M(T_0^* \leq t) \sim \mathbb{P}^*_{\overline{M-1}}(T_0^* \leq t) \tag{2.20}$$

asymptotically for M large. The statement then follows from the result in Proposition 2.4.6. □

As mentioned at the beginning of this section, the asymptotic equivalence between $\mathbb{P}_{\underline{1}}(T_M \leq t)$ and $\mathbb{P}^*_M(T_0^* \leq t)$ provides insight on why the former follows an extreme value type. Here, we stress again that the interpretation of T_0^* like a maximum of i.i.d. random variables is a direct consequence of the branching property, and can also be noticed looking at equation (1.7). From the same equation we also see that the distribution of the time to extinction for the progenies of each of the initial M individuals of Z_t^* is equal to

$$\mathbb{P}^*_1(T_0^* \leq t) = \frac{q\left(1-e^{\lambda t}\right)}{q-e^{\lambda t}}\,.$$

To check that such a distribution belongs to the Gumbel domain of attraction, we refer for example to Theorem 1.6.2 in [121].

Critical case

We now apply the same reversal argument described in the previous section to the critical case. Let Z_t be a critical branching birth-death process with rate α. Following the exact same steps as before, we find that the distribution of the time

to extinction for the reversed process Z_t^* starting with M individuals can again be expressed as

$$\mathbb{P}_M^*(T_0^* \leq t) = \mathbb{P}_M^*(R_M^* + X_0^* \leq t)\,, \tag{2.21}$$

where R_M^* is the last exit time from state M and X_0^* is a random variable whose distribution coincides with $\mathbb{P}_{\underline{1}}(T_M \leq t)$. However, for a critical branching birth-death process the last exit time R_M^* does not go to zero in probability as $M \to \infty$. In order to check this, let us first compute the asymptotic distribution of T_0^* conditional on M initial individuals. This follows again from equation (1.7) and is given by

$$\mathbb{P}_M^*(T_0^* \leq t) = \left(\frac{\alpha t}{\alpha t + 1}\right)^M .$$

By scaling T_0^* with M/α we then find

$$\mathbb{P}_M^*\left(\frac{\alpha {T_0}^*}{M} \leq t\right) = \mathbb{P}_M^*\left(T_0^* \leq \frac{Mt}{\alpha}\right) = \left(1 - \frac{1}{Mt+1}\right)^M \longrightarrow e^{-\frac{1}{t}}$$

as $M \to \infty$, which tells us that T_0^* belongs to the Fréchet domain of attraction (see Section 1.2.2). Now, the distribution of T_0^* conditioned on M initial individual and its limit discussed above both have infinite mean [20, 117]. On the other hand, the exact and asymptotic distribution of T_M have finite expectation (see Proposition 2.3.5), which implies that $\mathbb{E}_M^*[R_M^*] = \infty$ for every M. Therefore, the last return time R_M^* does not tend to zero in probability. As a consequence, in the critical case the reversed argument we presented does not relate the asymptotic behaviours of $P_{\underline{1}}(T_M \leq t)$ with distributions of an extreme value type.

2.5 Summary of results

In this chapter we have reviewed results concerning the probability distribution of hitting times in branching birth-death processes. For general birth-death processes with a reflecting boundary at the lower end of the state space, the exact form of this distribution is provided by a famous theorem from Karlin and McGregor. Such a theorem characterizes the distribution of the hitting time to M in terms of the M non-zero eigenvalues of the birth-death process infinitesimal generator. For branching birth-death processes, this characterization can be applied by conditioning on non-extinction, since in this case state 1 becomes a reflecting boundary.

The result from Karlin and McGregor is closely related to properties of the orthogonal polynomials associated to a birth-death process. In particular, the eigenvalues of the underlying infinitesimal generator can be expressed in terms of the zeros of these polynomials. For critical branching processes, the associated orthogonal polynomials are Laguerre ones with index 1, while for noncritical processes they do not correspond to a known class.

For the biological applications that motivate this work, the asymptotic distributions of hitting times to a large size are of great interest. Using the results for finite size, in the critical case we derived these distributions by exploiting the asymptotic limit of the zeros of Laguerre polynomials. In the noncritical case we had to use a different approach based on a classical martingale limit for branching processes. This argument showed that the hitting time to M in noncritical branching processes asymptotically follow a Gumbel distribution as M gets large.

The last asymptotic limit highlights a non-trivial connection with extreme value distributions, which we explored through reversed birth-death processes. We showed that as M gets large, the time taken by a supercritical branching process to go from size 1 to M is asymptotically distributed as the time to extinction for a process starting with $M-1$ individuals and with rates swapped. The latter time is the maximum of the times to extinction for the progenies of each of the initial individuals, and its distribution thus converges to an extreme value type. The same reasoning does not apply to the critical process.

The results presented in this and in the previous chapter provide us with the necessary tools to analyze our mathematical descriptions of cancer recurrence and biomarker shedding. In particular, the asymptotic distributions of hitting times in supercritical branching birth-death processes are fundamental to our model for the time of cancer relapse that we will introduce in the next chapter.

Chapter 3

Cancer recurrence times from a branching process model

3.1 Introduction

Metastases develop as cancer cells disseminate from a primary tumor and establish new malignant lesions in the surrounding tissue or at other sites [168]. However, the full process of metastasis formation is much more complex and many related aspects are not yet fully understood. In particular, it is still unclear whether metastases are initiated during early or late stages of carcinogenesis (see e.g. [140, 81, 158]). These details, however, affect the chances of a patient presenting detectable or undetectable metastases at diagnosis, which in turn influences treatment strategies and prognosis. For these reasons, different authors (see e.g. [135, 77] and the references therein) have proposed mathematical models to improve our understanding of the dynamics of metastasis formation.

Metastases frequently arise in cancer patients, and their occurrence greatly diminishes the chances of effective treatment. In fact, even when a therapy is initially successful, metastases often lead to relapse and are responsible for an estimated 90% of cancer related deaths [36]. Despite this common disease progression, reliable predictions for cancer recurrence rates and times are still lacking [189].

In Section 1.1 we have mentioned some of the many traits of tumor development that have been described through the Luria-Delbrück model and its generalizations. Around the 1950s, several researchers started to employ these mathematical frameworks to study temporal features of cancer evolution. In this con-

text, Armitage and Doll were the first to propose a stochastic model for the time to tumor onset [17]. A few decades later authors began to investigate stochastic models of tumor latency time. In particular, these works led to mathematical descriptions of optimal schedules of cancer surveillance [80, 79], cure rates [190] and cancer recurrence [211]. While this literature is based on different definitions of the random time to cancer relapse, the end point for all these random variables is the time to occurrence of a new tumor, and they are thus studied in the context of survival analysis. An excellent review of these models is provided in the book by Yakovlev and Tsodikov [212].

In this chapter we build a model for cancer recurrence by joining these two approaches. In particular, we consider a deterministically growing tumor seeding metastases at a rate depending on its size [113], and model the evolution of each metastasis (or clone) as independent birth-death branching processes. A similar setup was used by Lea and Coulson to mimic mutations occurring in a growing bacterial population [120]. In our model though we interpret these mutation events (from wild-type cells to mutants) as metastasis initiation events. The distribution of mutant close sizes was studied with an exponentially growing wild-type population [106] and with more general wild-type growth function [142]. Kendall [108] also allowed the wild-type population to grow stochastically, but this extension left the mutant behavior unchanged for small initiation (mutation) rates [110, 37]. Hence in our model we describe the size of the primary tumor as a deterministic function (focussing on exponential and logistic growth as examples), while allow the seeded metastases to grow according to birth-death branching processes. Within this framework we study the time to cancer relapse, defined as the interval between the primary onset and the first time that any of the metastases reaches a fixed detectable size. Similar characterizations are employed in the threshold models described in [212, 211].

Our mathematical model for the time to cancer recurrence relies on the use of a deterministic function and of branching birth-death processes to describe the growth of a primary tumor and metastases, respectively. To justify the former assumption, let us anticipate that metastases typically arise when a primary tumor is relatively large. Hence, even if we used a branching birth-death process to model also the evolution of a primary tumor, by the time of the first metastasis initiation this would already express an almost perfect exponential growth. On the other hand, fluctuations around small primary tumor sizes could affect the times of metastases initiation and relapse, measured from the primary onset.

However, since the time of this onset cannot be observed in practice, cancer recurrence is usually measured from the day of diagnosis or surgical resection of the primary, when again the tumor is already large. Hence, results computed from an estimated time of surgery with this fully stochastic framework would almost coincide with the predictions provided by our model with a deterministic exponential primary growth. The employment of branching processes to model the evolution of metastases is based instead on the assumption that competition for resources among cancer cells will likely occur only once the tumoral mass is large.

The rest of this chapter is organized as follows: In Section 3.2 we present in detail our mathematical model of metastases initiation and growth, and derive an explicit formula for the probability distribution of the time to relapse.

In Section 3.3 we include into the model the resection of the primary tumor at a given time. This allows us to distinguish between synchronous and metachronous metastases and to study the relapse time distribution conditioned on different events of clinical interest.

In Section 3.4 we report parameter estimates for five different cancer types (namely breast, colorectal, headneck, lung and prostate) and compare the corresponding results yielded by our model with data collected from clinical literature.

Conclusions are presented in Section 3.5, while supplementary information are shown in Appendix B.

3.2 Metastasis seeding and growth

Our mathematical characterization of the time to cancer recurrence is based on a stochastic model of metastasis formation. Here we first present the fundamental assumptions and features of this model, and then use them to derive the probability distribution of the time to relapse.

3.2.1 Model setup

We model the number of cells in the primary tumor as a deterministic function of time $n(t)$. The tumor initiates metastases at rate $\nu n(t)$, where ν is constant. Here we implicitly assume that all tumor cells can metastasize at the same rate. Since we make no assumptions on $n(t)$, one can define initiation at rate $\nu n(t)^\gamma$ to model scenarios where only a fraction of the primary tumor can metastasize, for example only the cells near its surface or close to blood vessels (see e.g. [77]). The initiated metastases are then modelled as independent branching birth-death

processes (see Section 1.2.1), all with the same birth rate α and death rate β. We assume that they are supercritical, that is they have a positive net growth rate $\lambda = \alpha - \beta > 0$ and that cancer cells from metastases do not have the ability to metastasize further [90].

Under these assumptions each metastasis will eventually go extinct with probability $q = \beta/\alpha < 1$. The surviving ones instead grow unboundedly and will reach any given size [20]. Let M be a fixed number of cells representing the minimal detectable size of a cancerous lesion. Here we aim to describe the time to cancer recurrence, defined as the first time τ that any of the surviving metastases reaches the detectable size M.

The minimal detectable size M is typically very large, with estimates larger than 10^6 (see Section 3.4.1). As the probability that a large supercritical population goes extinct is negligibly small, we assume that each metastasis survives if it reaches M. Then, due to the splitting property of Poisson processes, the surviving metastases that eventually reach the detectable size are initiated as a non-homogeneous Poisson process $(K_t)_{t\geq 0}$ with rate $\nu(1-q)n(t)$ (see Section 1.2.3). Here K_t denotes the number of metastases initiated by t, conditioned on survival. The expected number of established metastases at time t is thus

$$a_t = \mathbb{E}[K_t] = \nu(1-q)\int_0^t n(s)ds$$

and the probability that at least one is present at t is equal to

$$\mathbb{P}(K_t \geq 1) = 1 - e^{-a_t}\,. \tag{3.1}$$

Surviving metastases are initiated at times $\sigma_i := \inf\{t \geq 0 : K_t = i\}$ and are described by i.i.d. birth-death processes $(S_i(s))_{s\geq 0}$, where $S_i(s)$ is the number of cells in the i-th metastasis at time s after its establishment. In particular, we have $S_i(0) = 1$ for every i. For each of these processes we can then define $\Theta_i := \inf\{s \geq 0 : S_i(s) = M\}$ as the time needed by the i-th established metastasis to grow to the detectable size M, counting again from its initiation. Since the processes $S_i(s)$ are independent, the hitting times Θ_i are also independent and identically distributed. As shown in Theorem 2.4.1, for large M these hitting times asymptotically follows a Gumbel distribution. Hence, we have

$$\mathbb{P}(\Theta_i \leq t \mid \Omega_\infty^{(i)}) \sim G(t) = e^{-(1-q)Me^{-\lambda t}}\,, \tag{3.2}$$

where $\Omega_\infty^{(i)}$ denotes the eventual survival for the ith metastasis. Notice that hereafter we will always denote $G(t) = G(\frac{1}{\lambda}\log[(1-q)M], \frac{1}{\lambda}; t)$, where the distribution function $G(a, b; t)$ of the Gumbel for the maximum is defined by equation (1.20).

3.2.2 Time to reach detectable size

Given the definitions in the previous section, we have that the i-th metastasis reaches the detectable size at time $\tau_i := \sigma_i + \Theta_i$, measured from primary onset. Metastases are initiated at time s at rate $\nu(1-q)n(s)$ and then reach the detectable size before t with probability $G(t-s)$. Hence, the thinning property of Poisson processes (see equation (1.22)) yields that metastases which become detectable by time t are initiated at s at rate $\nu(1-q)n(s)G(t-s)$. The number S_t of such metastases established by t is thus a Poisson random variable with mean

$$b_t = \mathbb{E}[S_t] = \nu(1-q)\int_0^t n(s)G(t-s)ds\,. \tag{3.3}$$

The relapse time is defined as the first time any metastasis reach the detectable size, $\tau := \min_i\{\tau_i\}$. Hence, τ is smaller than t if by that time at least one metastasis that becomes detectable before t is initiated, and so

$$\mathbb{P}(\tau \le t) = \mathbb{P}(S_t \ge 1) = 1 - e^{-b_t}\,. \tag{3.4}$$

In the large detectable size M limit, the relapse time distribution converges to a simpler form (see Appendix B.1)

$$\tau - \frac{1}{\lambda}\log M \xrightarrow{d} \bar{\tau}\,,$$

where the random variable $\bar{\tau}$ is distributed as

$$\mathbb{P}(\bar{\tau} \le t) = 1 - \exp\left(-\nu(1-q)\int_{-\infty}^t n(t-s)e^{-(1-q)e^{-\lambda s}}ds\right)\,. \tag{3.5}$$

Hence for large M the relapse time decomposes as $\tau \approx \frac{1}{\lambda}\log M + \bar{\tau}$ into a deterministic part which depends only on λ and M, and a random fluctuation described by $\bar{\tau}$. This decomposition also allows us to estimate the expected value of the relapse time as $\mathbb{E}[\tau] \sim \frac{1}{\lambda}\log M + C$ where the constant $C = \mathbb{E}[\bar{\tau}]$ can be obtained from equation (3.5).

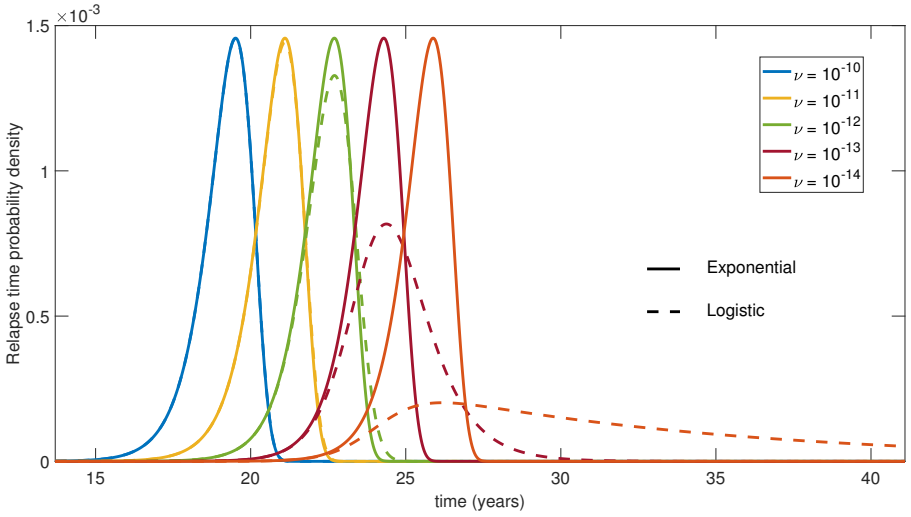

Figure 3.1: ***Relapse time densities for logistic and exponential primary growths and different initiation rates.*** *The probability densities shown,* $f_\tau(t) = \frac{d}{dt}\,\mathbb{P}(\tau \leq t)$*, are computed from equation* (3.4) *with* $\nu = 10^{-10}, 10^{-11}, 10^{-12}, 10^{-13}, 10^{-14}$ *from left to right. Using parameter estimates for colorectal cancer (see Table 3.2), the logistic densities (dashed lines) converge to the corresponding exponential ones as the initiation rate increases. Furthermore, in the exponential case and for all the above values of* ν*, the densities derived from equation* (3.4) *and their approximation obtained from equation* (3.6) *are indistinguishable.*

3.2.3 Exponential population growth

Two commonly employed primary growth functions are the exponential and logistic ones. These are given by $n(t) = e^{\delta t}$ and $n(t) = \frac{Ke^{\delta t}}{K+e^{\delta t}-1}$, respectively, where δ denotes the primary tumor net growth rate and K a carrying capacity. Relapse time densities for these two growth types and different initiation rates are shown in Figure 3.1. Here we observe that as ν increases, the logistic distributions converge to the exponential ones (see Appendix B.2). Moreover, for all our parameter estimates our model predicts the same results with these two growth types. The reason is that the metastases determining the time to relapse are initiated during the early phase of tumor evolution which is almost exponential even for a logistic growth. Therefore, from now on we will focus on the results for an exponential primary growth. Also notice that if only a fraction $n(t)^\gamma$ of the primary tumor cells can metastasize, for $n(t) = e^{\delta t}$ this would only affect the primary net growth rate.

Since the initiation rate ν is much smaller than all other parameters, here we study in detail the most relevant case, that is the small ν limit for an exponentially growing tumor. The deterministic part of the relapse time remains $\frac{1}{\lambda}\log M$, but

interestingly the fluctuations $\bar{\tau}$ are distributed as

$$\bar{\tau} \sim \text{Gumb}_{\min}\left(-\frac{1}{\delta}\log\frac{\nu(1-q)^{1-\frac{\delta}{\lambda}}\Gamma\left(\frac{\delta}{\lambda}\right)}{\lambda}, -\frac{1}{\delta}\right). \tag{3.6}$$

The first parameter of this Gumbel distribution is proportional to $\log\nu$, which explains the equal spacing between the densities in Figure 3.1 for logarithmically-spaced values of the initiation rate. Also notice that such densities are left skewed, as it is expected from the Gumbel for the minimum. On the other hand, the Gumbel for the maximum - which describes the fluctuations of the time to detection starting from a single initial cell - is right skewed (see Section 1.2.2). How one distribution changes to the other can be observed in Figure 3.3.

In the parameter regime considered here the mean relapse time is approximately given by

$$\mathbb{E}[\tau] \approx \frac{1}{\lambda}\log M + \frac{1}{\delta}\log\frac{\delta}{\nu} + C\,, \quad C = -\frac{1}{\delta}\left(\log\frac{\delta(1-q)^{1-\frac{\delta}{\lambda}}\Gamma\left(\frac{\delta}{\lambda}\right)}{\lambda} + \gamma_E\right). \tag{3.7}$$

As shown by Figure 3.2, this expression fits simulations even for relatively large values of ν and small values of M.

Equation (3.7) highlights a simple dependence of the mean relapse time $E[\tau]$ on M and ν. In Appendix B.2 we also compute the mean time to detectability of the first established metastasis, $\mathbb{E}[\tau_1]$, where $\tau_1 = \sigma_1 + \Theta_1$ is equal to the sum of the first initiation time and the hitting time to M. Interestingly $\mathbb{E}[\tau_1]$ has the same M and ν dependence shown in equation (3.7), but the constant term $\tilde{C}$ is different. For example, using the parameter estimates for colorectal cancer (see Table 3.2) we find $C \approx 250$ and $\tilde{C} \approx 309$. The reason for this difference is that even in the small ν - large M limit, later established metastases can outrun the earlier ones in reaching M first.

3.3 Primary tumor resection

Surgery is still the most common and effective type of treatment for solid tumors, although often used in combination with other kind of therapies (see e.g. [28]). However, how the time of resection affects prognosis, and in particular the estimation of the time to relapse, is still unclear. In order to investigate this question in a theoretical framework, we now embed surgery in our model and study how it changes the distribution of the time τ to relapse.

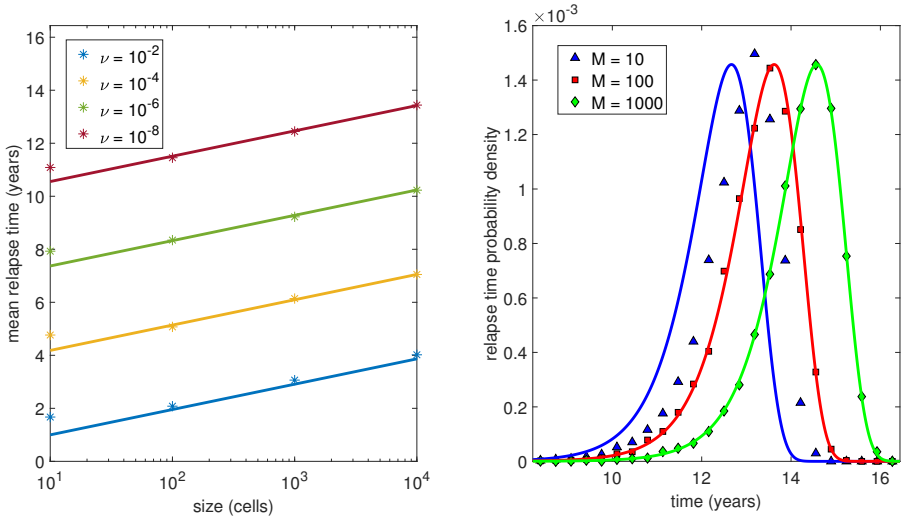

Figure 3.2: ***Relapse time distribution for exponential primary growth in the small initiation rate - large detectable size limit.*** *The figure is based on the parameter estimates for colorectal cancer - see Table 3.2. On the left, each starred dot denotes the mean of* 1000 *simulations, while lines represent the theoretical expectation given by equation* (3.7)*. These match the simulated means almost perfectly for most* ν *values, as the fit becomes poor only for* $\nu = 10^{-2}$ *or greater. On the right, the relapse time densities derived from equation* (3.6) *yield a bad approximation only for very small values of* M*, as the simulated data (*10000 *simulations per curve) are matched for* $M = 100$ *or higher.*

3.3.1 Relapse time with resection

Let us assume that at a given moment after detection a primary solid tumor is surgically removed. This event can be mathematically implemented in our model by considering a resection time T such that $n(t) \equiv 0$ for $t \geq T$. In particular, this implies that after T no metastases can be initiated. The number of metastases already established at resection is equal to K_T, and their size distribution is given in [142]. The distribution of the time τ to relapse can then be expressed exactly as in equation (3.4), however here τ is not a proper random variable. In fact, as $\int_0^\infty n(s)ds = \int_0^T n(s)ds < \infty$, there is a positive probability that no metastasis will ever occur (notice that from this point of view our framework can be seen as a cure model - see e.g. [152]) and in this case we set $\tau = \infty$. The distribution of the relapse time conditioned on at least one metastasis established by resection is simply

$$\mathbb{P}(\tau \leq t \mid K_T \geq 1) = \frac{\mathbb{P}(\tau \leq t)}{\mathbb{P}(K_T \geq 1)} = \frac{1 - e^{-b_t}}{1 - e^{-a_T}}\,. \tag{3.8}$$

This conditional distribution for different resection times is depicted in Figure 3.3. In this and following figures, the resection time is shown at the bottom of the figure, and the corresponding resection size $N = e^{\delta T}$ is shown on the top. As

$T \to 0$ all metastases have to be initiated close to time zero, so the relapse time becomes the time to reach size M from a single cell, which has the Gumbel distribution for the maximum given by equation (3.2). If we then increase the resection time, the conditional densities shift to the right by the same amount. Finally, as $T \to \infty$ the relapse time distribution converges to the case without resection

$$\mathbb{P}(\tau \leq t \mid K_T \geq 1) \to \mathbb{P}(\tau \leq t)\,.$$

The fluctuations for the unconditional distribution follow a Gumbel type for the minimum, as per equation (3.6). Hence, as time increases, the relapse time distribution turns from a right-skewed Gumbel to a left-skewed Gumbel.

Note that the densities in Figure 3.3 become indistinguishable from the large time limit as $\mathbb{P}(K_T \geq 1)$ approaches one. The reason is that by this time metastases have probably already been initiated and one of the early established ones is likely to relapse first. This suggests that only early enough resection times change the behaviour of the model. For example in the case of colorectal cancer, according to Figure 3.3, only resections of tumors smaller than 10^9 cells affect the time to recurrence.

Right skewed densities are often chosen to fit probability distributions arising in survival analysis. This is due to the fact that most survival data suffer from right censoring [9], where only a lower bound is known for data points. Looking at the densities in Figure 3.3, though, we can see both left and right skewed distributions. While a few survival datasets are negatively skewed [133], cancer relapse times are typically right censored as a consequence of limited follow-up and patients decease before relapse (see e.g. [174]). However, our model does not take into account any of these events. Furthermore [78] recently proposed a model for the estimation of screening times for colorectal cancer based on the observation that some datasets suffer from left censoring as well.

3.3.2 Metastasis classification

If the resection is successful and the primary tumor is completely removed, the therapy can still fail due to the formation of metastases. For this reason, it is common practice to start looking for detectable metastases several weeks before the surgery. In this section we thus want to characterize the metastases which are detectable at a given time and those which are not.

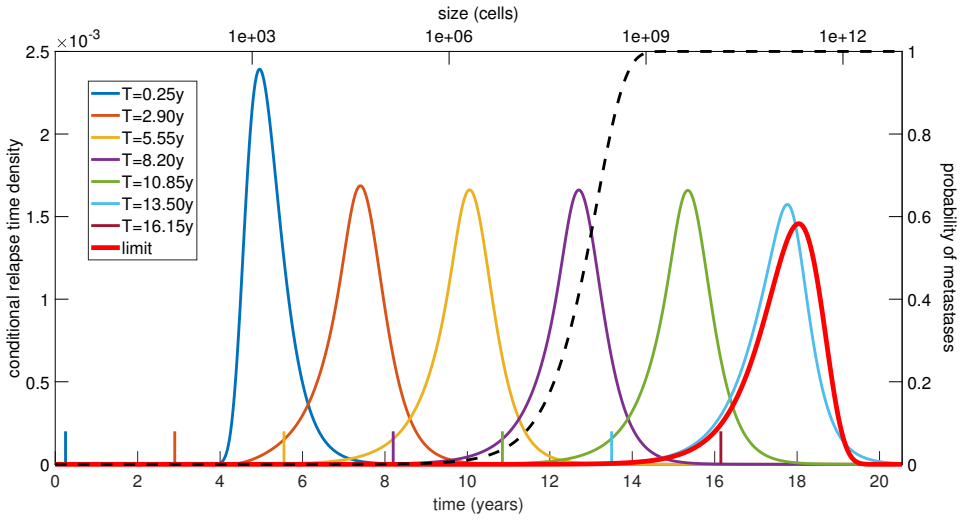

Figure 3.3: ***Relapse time densities conditioned on at least one metastasis initiated by the time of resection.*** *For different values of the resection time T, marked with ticks of corresponding colors, the densities $f_\tau(t \mid K_T \geq 1)$ are computed by differentiating equation* (3.8)*. As T becomes larger, the probability of metastases established before resection (see equation* (3.1)*) increases and the conditional relapse time densities converge to the red limit one. Here we have used parameter estimates for colorectal cancer (see Table 3.2), $n(t) = e^{\delta t}$ and 7 equally spaced resection times between $0.25y$ and $16.15y$. The curves for $T > 15y$ look identical to the limit density.*

In general the metastasizing process $(K_t)_{t\geq 0}$ can be split into two independent Poisson processes $(S_t)_{t\geq 0}$ and $(M_t)_{t\geq 0}$ describing the initiation of metastases which reach size M before or after t, respectively. Following the same argument we used at the beginning of Section 3.2 we see that a mean number b_t of metastases of the former type are initiated by t. Therefore we get

$$S_t \sim \text{Pois}\,(b_t)\,, \qquad M_t \sim \text{Pois}\,(c_t)\,,$$

where $c_t = a_t - b_t$. In particular, we have that the events $\{\tau > t\}$ and $\{S_t = 0\}$ are equivalent. We also stress that the definitions above naturally extend to the case of a primary resection, by simply redefining $n(t)$ to be zero after the resection time T.

Now, despite an ongoing discussion on the following nomenclature (see e.g. [4]), in the rest of the paper we will call a metastasis synchronous if it reaches the detectable size M before or up to the time of resection, and metachronous otherwise (hereby the choice of notation S_t and M_t). These characterizations immediately allow to estimate the probability of clinically relevant events. For

example, the probability of no synchronous metastases is equal to

$$\mathbb{P}(S_T = 0) = \mathbb{P}(\tau > T) = e^{-b_T} . \tag{3.9}$$

Also, under this condition, relapse is not certain: the probability that at least one metastasis was initiated given that there are no visible ones at resection is

$$P(K_T \geq 1 \mid S_T = 0) = \mathbb{P}(M_T \geq 1) = 1 - e^{-c_T} , \tag{3.10}$$

since S_T and M_T are independent. In next section we will study the above and related quantities in greater detail.

3.4 Comparison to data

In this section we compare the predictions provided by our model with clinical data collected for different cancer types. To this purpose, we first need to estimate the parameter values for each of these cancer types.

3.4.1 Parameter estimation

The net growth rates of the primary and metastatic tumors, δ and λ, are inferred from the corresponding tumor volume doubling times (denoted DT_{pt} and DT_m, respectively) as

$$\delta = \log 2/DT_{pt} , \quad \lambda = \log 2/DT_m .$$

These times have been studied by many authors, starting from the influential papers of [43, 171, 177]. Many authors still refer to these early works, although in some cases more recent estimates are available. Colorectal, breast and lung cancers are the most frequently studied. Furthermore, more papers focus on primary doubling times than on metastatic ones. Similarly, the birth rate α is derived from the potential doubling time T_{pot}, defined as the time between cell divisions in the absence of cell death [97]. In this case we simply use the estimation

$$\alpha = 1/T_{pot} .$$

As for the primary tumor size N at resection, many studies report data on the primary maximum diameter, allowing for ellipsoidal forms. However, given the relatively small tumor volume and the wide interpatient variability, we assume a spherical shape and estimate d_{pt} from the corresponding typical range. By also assuming 10^9 cells per cm^3, the primary size at resection (expressed in number of cells) is thus estimated as $N = \frac{1}{6}\pi d_{pt}^3 10^9$.

Table 3.1 summarizes typical ranges of these quantities for five different cancer types, together with our estimates and corresponding literature references. Difficulties in distinguishing between primary and secondary tumors or in tracking down the primary origin of a metastatic cancer could in principle affect some of these data, but the wide range and multiple references reported reduce the potential impact of this effect.

Table 3.1: ***Typical ranges of tumor volume doubling times and tumor diameter at resection for breast, colorectal, headneck, lung and prostate cancer.***

Cancer type	Parameter	Typical range	Estimate	References
Breast	DT_{pt} (days)	$103 - 353$	210	[197, 118, 151, 167, 71, 218]
	DT_m (days)	$85 - 199$	105	[119, 69]
	T_{pot} (days)	$8 - 35$	15	[82, 21]
	d_{pt} (cm)	$1.4 - 3$	2.5cm	[216, 122, 49]
Colorectal	DT_{pt} (days)	$130 - 438$	175	[27, 183, 39]
	DT_m (days)	$45 - 155$	105	[64, 69, 185, 187]
	T_{pot} (days)	$3 - 4$	4	[209, 21]
	d_{pt} (cm)	$3.5 - 5.1$	4.5cm	[27, 116, 39, 52]
Headneck	DT_{pt} (days)	$15 - 256$	84	[198, 94]
	DT_m (days)	$9.5 - 320$	56	[72, 192]
	T_{pot} (days)	$1 - 14$	4	[209, 217]
	d_{pt} (cm)	$1.3 - 4$	2.8cm	[138, 130]
Lung	DT_{pt} (days)	$22 - 269$	168	[109, 15, 69, 50, 86]
	DT_m (days)	$32 - 98$	56	[177, 215]
	T_{pot} (days)	$2 - 17.5$	2.5	[109, 68]
	d_{pt} (cm)	$1.7 - 4.1$	2cm	[22, 178, 50]
Prostate	DT_{pt} (days)	$36 - 1080$	392	[46, 206, 220]
	DT_m (days)	$29 - 213$	98	[24, 220]
	T_{pot} (days)	$15.2 - 97.8$	34	[83, 206]
	d_{pt} (cm)	$0.1 - 2.9$	1.2cm	[159, 96]

Notice that by estimating the rates λ and α we also infer values for the death rate $\beta = \alpha - \lambda$ and the extinction probability $q = 1 - \lambda/\alpha$. These estimates are based on the assumption of an exponential primary growth, which can be relaxed as in [184]. For the two remaining parameters, namely the initiation rate ν and the minimal detectable size of a metastasis M, we use common estimates across different cancer types. Various studies report a lowest detectable tumor diameter

of 0.2cm for different cancer types (see e.g. [173, 70, 200]), corresponding to $M \approx 4.19 \times 10^6$ cells. Moreover, several papers argue that the first metastases are likely to establish long before the detection of the primary tumor (see for example [69] and the references therein). In particular, the review of the progression model for metastases formation in [113] reports that dissemination starts when the primary diameter is between 0.1 and 0.4cm. Here, we thus consider the primary tumor size at the expected time of the first metastasis initiation and estimate it to be $e^{\delta \mathbb{E}[\sigma_1]} = 10^8$ cells, corresponding to a diameter of about 0.58cm. Hence, referring to the results in Appendix B.2 we use

$$\nu \approx \frac{\delta e^{-\gamma_E}}{1-q} e^{-\delta \mathbb{E}[\sigma_1]} .$$

Finally, the carrying capacity for the logistic primary growth is set to $K = 10^{12}$ [113, 38]. Overall, we thus found estimates for the following input vector

$$(DT_{pt}, DT_m, T_{pot}, d_{pt}, d_m, e^{\delta \mathbb{E}[\sigma_1]})$$

and used them as described above to derive values for our model parameters, i.e.

$$(\delta, \lambda, \nu, q, N, M) .$$

Such estimates are summarized in Table 3.2.

Table 3.2: ***Estimates for the model parameters.*** *The primary tumor and metastases net growth rates δ and λ and the initiation rate ν are measured in cells per day. The typical resection and detection size, N and M respectively, are expressed instead in number of cells. Finally, the estimated extinction probability q is a pure number.*

	Breast	**Colorectal**	**Headneck**	**Lung**	**Prostate**
δ	0.0033	0.0040	0.0083	0.0041	0.0018
λ	0.0066	0.0066	0.0124	0.0124	0.0071
ν	1.87×10^{-10}	8.42×10^{-10}	9.36×10^{-10}	7.49×10^{-10}	4.13×10^{-11}
q	0.9010	0.9736	0.9505	0.9691	0.7595
N	8.18×10^9	4.77×10^{10}	1.15×10^{10}	4.19×10^9	9.05×10^8
M	4.19×10^6	4.19×10^6	4.19×10^6	4.19×10^6	4.19×10^6

To conclude this section, let us mention that by employing branching birth-death processes to model the evolution of metastases, we are implicitly assuming that the lifespans of tumor cells are exponentially distributed. This assumption was primarily made to keep the model simple and to be able to apply and test the results presented in Chapter 2. However, while the duration of cell cycle is

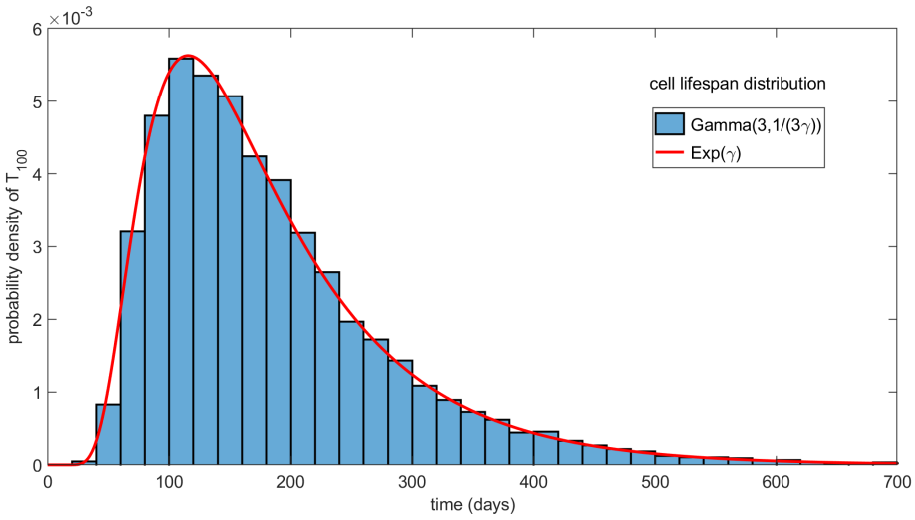

Figure 3.4: ***Hitting times with Gamma distributed cell cycle duration.*** *The figure shows a first passage time distribution for a stochastic process $\tilde{Z}_t$, defined as a process starting with one individual, jump probabilities equal to those of the branching birth-death process Z_t, and indivual lifetimes distributed according to a $\Gamma\left(3, \frac{1}{3\gamma}\right)$ distribution. In particular, the histogram shows the distribution of the first passage time to $M = 100$ obtained from 10^5 simulations of the process $\tilde{Z}_t$. The red curve illustrates instead the exact theoretical density of T_{100} for the branching-birth death process Z_t, derived from Theorem 2.3.1. Using our estimates values of α and β for colorectal cancer (see Table 3.2), and already for $M = 100$, the two models provide almost identical results.*

likely to follow more complicated distributions (see e.g. [137, 205]), we expect that the shape of these distributions should not affect significantly our results, provided that their mean and variance are approximate those of the exponential interevent times. An intuitive explanation for this feature is that if a branching birth-death process is sufficiently close to criticality - as it is the case for all our estimates in Table 3.2 - then it takes on average a long time and a significant number of jumps to reach even relatively small sizes, and differences in the shape of cell lifespan distributions get lost across such a long time. As an example, we have simulated the distribution of the first passage time to 100 cells for a stochastic process where each cell lifespan follows a $\Gamma\left(3, \frac{1}{3\gamma}\right)$ distribution. Such an assumption takes into consideration the three phases of a cell cycle described in [205], while the expected cycle duration is the same as for the branching birth-death model. These simulations are shown in Figure 3.4, and already for $M = 100$ they are perfectly matched by the theoretical density of T_{100} computed from Theorem 2.3.1. While this suggests that our predictions based on a branching process model are realistic, we aim to investigate further the effect of alternative cell lifespan distributions in a future study.

3.4.2 Model predictions

By employing the parameter estimates reported in Table 3.2 we can analyze the corresponding predictions of our model and compare them to clinical data.

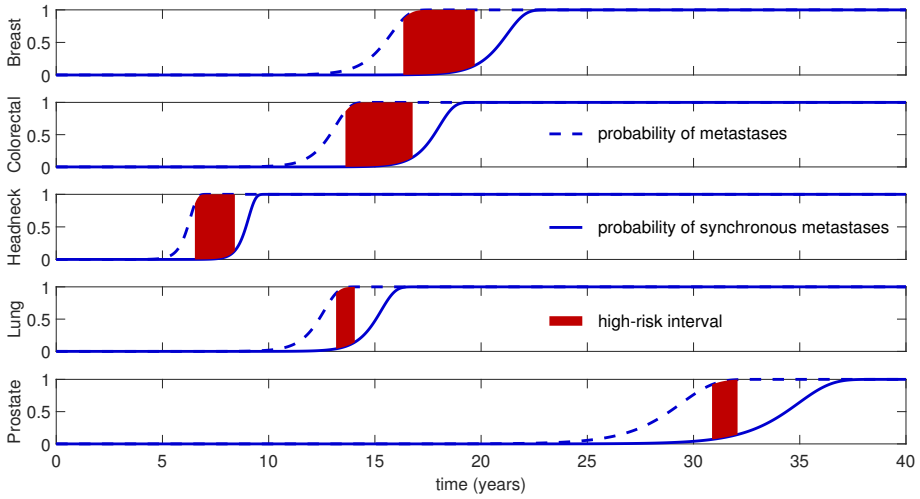

Figure 3.5: ***Probability of extant and synchronous metastases.*** *These probabilities* $\mathbb{P}(K_T \geq 1)$ *- dashed curve computed from equation* (3.1) *- and* $\mathbb{P}(S_T \geq 1)$ *- solid curve computed from equation* (3.4) *- are plotted as functions of the resection time* T *for five different cancer types. The primary tumor size at resection is* $N = e^{\delta T}$ *and thus depends on the primary net growth rate. These resection sizes are discussed Table 3.3. For each cancer type, the shaded areas highlight resection time intervals leading to a probability higher than* 85% *of established and all undetectable metastases. Using the parameter estimates from Table 3.2, the width of these intervals is* $3.41, 3.17, 1.92, 0.94, 1.19$ *years for breast, colorectal, headneck, lung and prostate cancer respectively.*

Let us start by the simplest predictions of the model, which are about the presence of synchronous and metachronous metastases. Figure 3.5 shows the probability of initiated metastasis by resection $\mathbb{P}(K_T \geq 1) = 1 - e^{-a_T}$ (equation (3.1)), and that of synchronous metastasis $\mathbb{P}(S_T \geq 1) = 1 - e^{-b_T}$ equation (3.4)) as functions of the resection time, for five different cancer types. Clearly, metastases establish much before any of them becomes visible. For all five cancer types considered, one or more metastases have likely been initiated by the time the primary tumor reaches about 8.2×10^8 cells (diameter 1.16cm). While this value is similar across different primary types (as a consequence of the parameters estimation procedure, see Section 3.4.1), the results for the probability of synchronous metastases vary widely. For breast, colorectal, headneck, lung and prostate cancer, Table 3.3 reports primary tumor sizes at which synchronous metastases might start to appear and are likely to be present, respectively (expressed both in terms of number of cells and tumor diameter). By comparing these values to typical

Table 3.3: ***Resection sizes of the primary tumor which yield low and high probability of synchronous metastases, respectively.*** *For each cancer type considered, these sizes are computed with the parameter values in Table 3.2 and expressed both in terms of number of cells, N, and tumor diameter, d.*

		Breast	**Colorectal**	**Headneck**	**Lung**	**Prostate**
$\mathbb{P}(S_T \geq 1) > 0.01$	N	1.32×10^9	2.13×10^9	7.03×10^9	1.03×10^8	6.27×10^7
	d	1.36	1.60	2.38	0.58	0.49
$\mathbb{P}(S_T \geq 1) > 0.99$	N	6.03×10^{11}	9.88×10^{11}	3.22×10^{12}	4.65×10^{10}	2.89×10^{10}
	d	10.48	12.36	18.32	4.46	3.81

resection sizes in Table 3.1, we find that detecting metastases at resection is very likely for lung and prostate cancer and rare for headneck primary tumors.

One of the most challenging scenarios for the development of an effective treatment is when there are only undetectable metastases present. In our framework this scenario corresponds to the event

$$U_T := \{K_T \geq 1, S_T = 0\},$$

which has probability (see equations (3.9) and (3.10))

$$\begin{aligned} \mathbb{P}(U_T) &= \mathbb{P}(M_T \geq 1, S_T = 0) = \mathbb{P}(M_T \geq 1)\mathbb{P}(S_T = 0) \\ &= e^{-b_T} - e^{-a_T} = \mathbb{P}(K_T \geq 1) - \mathbb{P}(S_T \geq 1). \end{aligned} \tag{3.11}$$

Because of the last identity, the probability of established and all metachronous metastases can be read out from Figure 3.5 as the difference of the two curves. There, the shaded areas highlight intervals of resection times yielding $\mathbb{P}(U_T) > 85\%$. These intervals, often referred to as high-risk period, are alarmingly wide, especially for breast, colorectal and headneck cancers. Furthermore, the estimated resection sizes given in Table 3.1 fall within or close to these ranges ($\mathbb{P}(U_T)$ equal to $93.87\%, 79.83\%, 98.35\%, 66.04\%$ and 85.85% for the five primary tumor types studied, respectively).

In order to check the robustness of this feature, we analyze it for a range of parameters values. In particular, Figure 3.6 shows the probability of U_T for different values of the primary net growth rate δ and of the initiation rate ν, focussing on the parameter estimates for colorectal cancer. The width of the high risk interval is constant with respect to ν, and shrinks only as the ratio

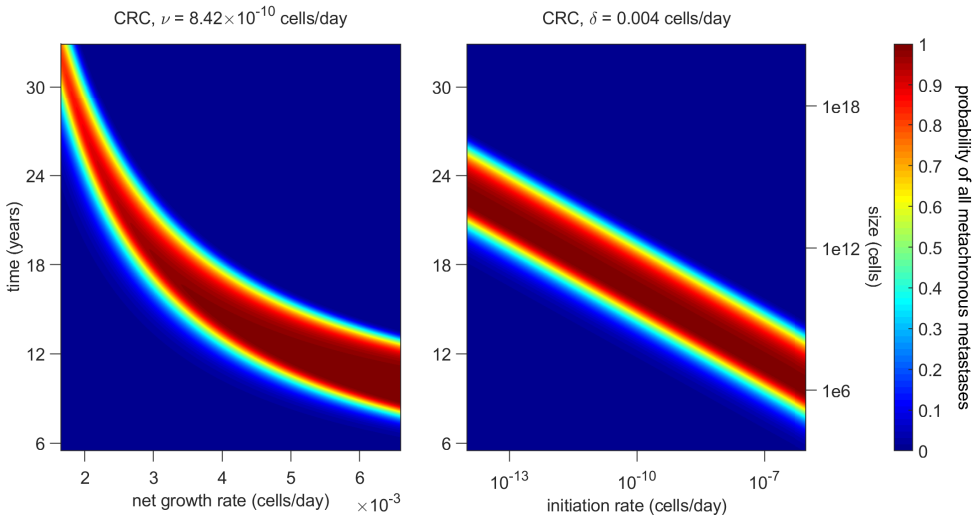

Figure 3.6: ***Probability of established and all metachronous metastases.*** *The probability* $\mathbb{P}(U_T)$ *- as given by equation* (3.11) *- is plotted as a function of* T *and* δ *(left panel) and* T *and* ν *(right panel). The parameter estimates used are those for colorectal cancer reported in Table 3.2. The plots show that the width of the high-risk interval - the range of resection times such that* $\mathbb{P}(U_T)$ *is high - stays roughly constant for most parameter values. This width (about* 3 *years) shrinks only for metastases growing significantly faster than the primary tumor that initiated them.*

between the primary and metastatic net growth rate becomes very small. The same qualitative behaviour can be obtained with the parameter estimates for the other cancer types. As most metastases grow up to two times faster than the primary tumor they originated from [113], our model suggests that for a wide choice of parameters there is a substantial range of resection sizes that lead to a high probability of established and all undetectable metastases.

Next, we ask how such a probability, $\mathbb{P}(U_T)$, influences the time to cancer recurrence. The conditional distribution of the relapse time τ becomes

$$\mathbb{P}(\tau \leq t \mid U_T) = \frac{e^{-b_T} - e^{-b_t}}{e^{-b_T} - e^{-a_T}},$$

for $t \geq T$. From this distribution we compute the expected relapse time after resection and conditioned on U_T, $\mathbb{E}[\tau - T \mid U_T]$. This expectation and the probability $\mathbb{P}(U_T)$ are plotted in Figure 3.7. We see that for resection sizes smaller than 10^8 cells the relapse occurs on average between 4 and 5 years after resection, otherwise independently of the primary size. For resection sizes around 10^8 cells undetectable metastases become likely, and $\mathbb{E}[\tau - T \mid U_T]$ starts to decrease with tumor size. Then, at about 19 years the probability of only undetectable metastases present and the conditional mean relapse time both approach zero.

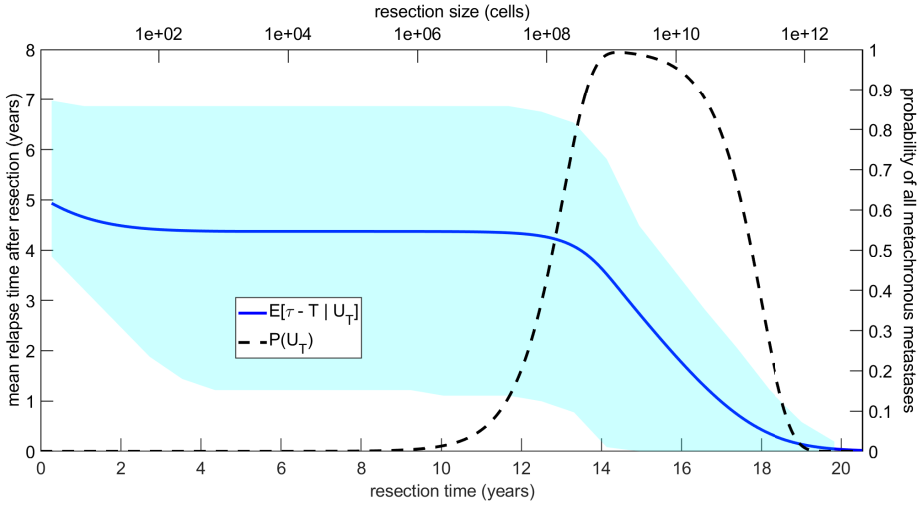

Figure 3.7: ***Expected relapse time measured from resection, conditioned on extant but all undetectable metastases.*** *The dashed line and the light blue shaded area show* $\mathbb{P}(U_T)$ *and how spread is the conditional relapse time distribution, respectively, while the solid blue curve represents* $\mathbb{E}[\tau - T \mid U_T]$*. The parameter estimates used are those for colorectal cancer reported in Table 3.2. For resection times close to zero this conditional expectation coincides with that of the Gumbel distribution given by equation* (2.14)*, at about* 5 *years. As* T *starts to increase* $\mathbb{E}[\tau - T \mid U_T]$ *reflects the convergence highlighted for Figure 3.3, first slightly decreasing and then staying constant around* 4.4 *years. Finally, when the resection time falls into the high-risk window, the expected relapse time drops to zero. This suggests that the bigger the primary tumor size is at resection, the faster relapse will occur.*

Using the values from Table 3.2 we then tested our model by computing the probability of synchronous metastases and the mean relapse time conditioned on established but all undetectable metastases. The predictions from our model, typical ranges and references for each cancer type considered are summarized in Table 3.4. Here, notice that our predictions for the mean relapse time fall on the lower end of the respective typical ranges. This is expected since we compute the time to recurrence τ based on the minimal detectable size M, while in practice metastases are often detected only at larger sizes.

In general, for different cancer types it is observed that metastases can grow up to 2 times faster than the primary tumor they originated from [113], although values as high as 4 has been proposed [123]. Our estimates fall within this range ($\delta/\lambda = 4$ for prostate cancer, 3 for lung and between 1.5 and 2 for the others). As per the time interval from primary onset to surgery, the typical range is $15 - 25$ years [97]. Here, the high variability in our estimates of DT_{pt} make T fall outside

Table 3.4: ***Typical ranges for the probability of synchronous metastases and the expected relapse time after resection, their predicted values from the model and literature references for each cancer type.*** $\mathbb{P}(S_T \geq 1)$ *is a pure number, while* $\mathbb{E}[\tau - T \mid U_T]$ *is measured in days.*

Cancer type	Output	Range clinical data	Theoretical prediction	References
Breast	$\mathbb{P}(S_T \geq 1)$	$5 - 10$	6.13	[12, 30, 213]
	$\mathbb{E}[\tau - T \mid U_T]$	$590 - 1022$	725	[111, 65, 145]
Colorectal	$\mathbb{P}(S_T \geq 1)$	$15 - 25$	20.17	[107, 149, 128, 56, 88, 213]
	$\mathbb{E}[\tau - T \mid U_T]$	$353 - 760$	356	[87, 143, 56, 179, 88]
Headneck	$\mathbb{P}(S_T \geq 1)$	$1 - 16.8$	1.65	[62, 93]
	$\mathbb{E}[\tau - T \mid U_T]$	$219 - 623$	435	[126, 55, 207]
Lung	$\mathbb{P}(S_T \geq 1)$	$30 - 55.39$	33.96	[213, 191]
	$\mathbb{E}[\tau - T \mid U_T]$	$210 - 602$	249	[7, 89, 60]
Prostate	$\mathbb{P}(S_T \geq 1)$	$10 - 34$	13.53	[115, 66, 10]
	$\mathbb{E}[\tau - T \mid U_T]$	$730 - 1131$	969	[29, 188]

that range for headneck ($T = 7.69$y), lung ($T = 14.71$y) and prostate ($T = 32$y) cancers, classifying the first two as fast growing tumors and the latter as a slow growing one. The singular features that the model predicts for prostate cancer are in accordance with clinical studies (see e.g. [170, 24]).

The last trait of cancer recurrence that we are going to examine is disease-free rates. These generally correspond to the survival function of the relapse time, $\mathbb{P}(\tau > t)$. However, following the previous discussion we will condition this probability on no synchronous metastases, obtaining

$$\mathbb{P}(\tau \leq t \mid S_T = 0) = 1 - e^{-(b_t - b_T)}\,, \tag{3.12}$$

for $t \geq T$. In this case we do not observe any convergence to the density without resection, because if $T \to 0$ then no metastasis can be initiated and if $T \to \infty$ the condition $S_T = 0$ pushes the relapse time to infinity. Here let us also stress that our model does not provide information on survival rates, as no modelling of the time to decease is incorporated. Furthermore, notice that $P(\tau > t)$ yields a good description of the disease-free rates in terms of metastases detectability, but not necessarily with respect to cancer symptomaticity.

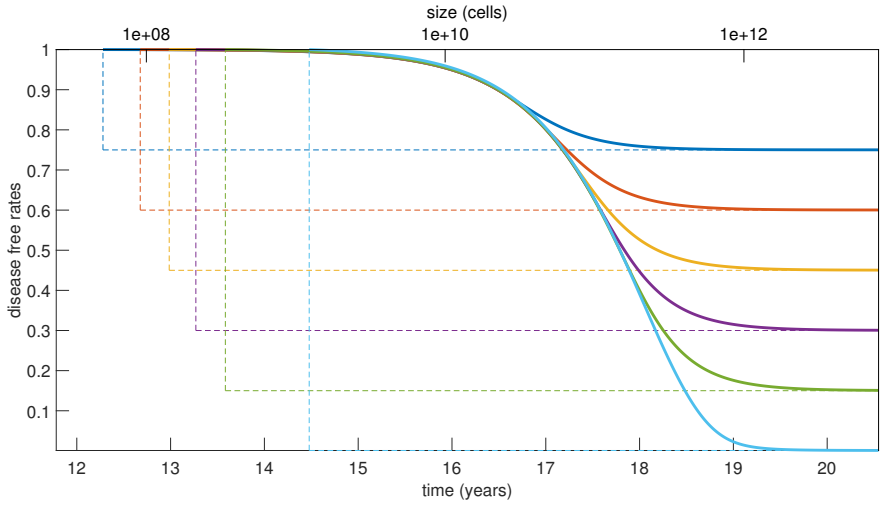

Figure 3.8: ***Disease-free curves for different resection times.*** *The earlier the primary tumor is resected the higher is the probability that no metastases will arise, or cure probability, represented by the value of the final plateaus. The resection times are chosen so that* $\mathbb{P}(K_T = 0) = 0.75, 0.6, 0.45, 0.3, 0.15, 0.001$ *respectively. With the parameter estimates for colorectal cancer (see Table 3.2) these times range from* 12.28 *to* 14.48 *years, corresponding to sizes between* 5.12×10^7 *and* 1.23×10^9 *cells (diameter* $0.46 - 1.33cm$*), respectively.*

The distribution $\mathbb{P}(\tau > t \mid \tau > T)$ for different resection times is shown in Figure 3.8, studying again the case of colorectal cancer. As we are not conditioning on at least one metastasis being initiated, there is always a positive probability that relapse will not occur, that is $\tau = \infty$. The resection times are thus chosen so to yield cure probabilities - $\mathbb{P}(K_T = 0)$, corresponding to the final plateaus - equal to $0.75, 0.6, 0.45, 0.3, 0.15$ and 0.001, respectively. These times span across a total range of about 2.2 years. Furthermore, excluding the latest resection time considered, the difference between two consecutive of these T values is between 0.28 and 0.4 years. Hence, our model suggests that delays of the order of months in the time of primary resection lead to a significant decrease in the cure probability.

3.5 Discussion

We introduced a model of metastasis formation where metastases are initiated at a time dependent rate, in the simplest case proportional to the size of a growing primary tumor. All initiated metastases then evolve as independent supercritical branching processes. Parameters of the model were estimated for five different cancer types from the clinical literature. We studied the relapse time τ, that is the earliest time when any of the metastases becomes detectable. We obtained

the distribution of τ for a general primary tumor growth and focussed in particular on logistic and exponential growth functions. For clinically relevant initiation rates the metastases which relapse first are typically initiated in the early phase of the primary tumor development, which is exponential for both growth functions considered. Hence the distribution of τ for exponential and logistic primary growths is practically identical unless the initiation rate is unrealistically small ($\nu \approx 10^{-13}$ or smaller) and we can thus exploit the much simpler formulas for the exponentially growing tumor.

We model the resection of the primary tumor by introducing a cut-off for the growth function $n(t)$. If metastases are likely already established at surgery, their time of relapse is not influenced by the resection timing. We categorized all metastases into synchronous and metachronous and computed corresponding occurrence probabilities. With our estimated parameters we found that the probability of synchronous metastases and the mean relapse time after resection falls in the typical clinical range for all five different cancer types we study.

A challenging scenario for treatment is that of patients with established but all undetectable metastases. For all five cancer types we considered, the probability of this event is high within an alarmingly large range of resection sizes. Unfortunately, the typical size of a resected tumor falls in or near this range for all cancer types. The width of such a high-risk range is stable for varying values of the initiation rate ν and the primary net growth rate δ. While conditioned on the presence of initiated but all undetectable metastases, later resection times lead to faster relapse after T. Relatively small delays in these resection times also cause significant decrease in the cure probability. Within our model, surgery only prevents recurrence if it is done before the onset of the first surviving metastases, and we provided estimates for the primary tumor sizes at this onset time.

The parameter estimates summarized in Table 3.2 yield realistic predictions for several quantities of clinical interest. Although in principle we can explore our model predictions across the whole range of parameters, this would often lead to unrealistic outcomes. In this sense the quantitative predictions of our model are quite sensitive to the parameter values, but we have been able to find a combination of parameters that yields realistic results. On the other hand, the qualitative features of our model are more robust to parameter changes, as demonstrated for example in Figure 3.6.

Metastases are seeded and establish colonies via a specific and complex process called metastatic cascade (for details see e.g. [146]). Since this is known to be a multi-stage process, some authors (see for example [53, 77, 76] and references therein) have described metastases initiation through two-type stochastic models, where a cell needs to gain the ability to metastasize before it can establish a new metastatic lesion. We did not choose that route for several reasons: (i) the precise details of how and when cells reach the ability to metastasize are not clear [97, 210], (ii) in our model we can think of $n(t)$ as the number of cells which can metastasize and so tailor the two approaches, and (iii) if we assume that an acquired metastatic ability lowers the primary net growth rate, a branching process model would predict the same exponential growth for the cells with this ability [37], and hence this would only change the estimate of the initiation rate in our model.

We did not include into our model a mechanism for metastasis seeding other metastasis, although this phenomena has been observed in clinical studies [75]. The main reason for this omission was the lack of reliable data for the estimation of the secondary seeding rate. By assuming the same primary and secondary seeding rates, however, we would expect metastases to initiate secondary ones when they reach around 10^8 cells, at which size they are already detectable. Hence, by considering this scenario our predictions for the time to cancer relapse would not change.

We aim to compare our model in the future to data where relapse times are given jointly with primary tumor sizes at resection. Tumor size is of course not the only relevant factor in predicting relapse times, so the model should be extended to involve other features like a measure of malignancy perhaps, as in [31]. Many of the parameters of the model can differ between patients, and also between each metastasis. Therefore, including a probability distribution for the parameters could also make our model more realistic, provided that such distributions can be estimated from data. Furthermore, as we stressed at the end of the parameter estimation section, we could consider alterative distributions for the cell cycle duration and quantify precisely their effects on the predictions of our model. Other possible extensions could include interactions among metastatic cells and among metastatic lesions, effects of the immune system, allowing metastases to seed other metastases, and providing an estimate for the fraction of cells which can metastasize, perhaps through modelling angiogenesis.

Chapter 4

Stochastic models of ctDNA shedding

4.1 Introduction

Among the results discussed in Chapter 3, our model of cancer recurrence suggested that early resection of a primary tumor decreases the probability of a patient developing metastases and thus improves cure rates. In general, it has been long known that the early detection of cancer can lead to better prognosis and lower the costs related to the management of the disease [58]. For this reason, significant research efforts have been dedicated to the enhancement of screening and diagnostic techniques.

However, some of the standard approaches currently available present limitations that are difficult to overcome. Modern imaging modalities based on positron emission or computed tomography and magnetic resonance scans can detect tumors as small as 7mm in diameter [57]. While such a lower detection limit is bound to decrease with the advancements of these technologies, it is hard to predict if and when they will allow to detect cancer before the onset of symptoms or the establishment of secondary tumors. Furthermore, once a tumor has been detected, tissue biopsies are often performed to determine some of its biological and genetic features. These procedures are often costly and invasive for the patient. Moreover, it might be impossible to collect samples from some tumors, and even when tissue biopsies are feasible they cannot be repeated over time.

Recently, the development of new screening tests based on the analysis of bodily fluids spurred new hope for early cancer detection [125, 84, 85]. These

tests, collectively referred to as liquid biopsies, search a patient blood sample for specific biomarkers whose concentration and genetic composition can provide information about the presence of a tumor and some of its attributes. Compared to solid biopsies, liquid biopsies are a minimally invasive test which can drastically reduce the impact and costs of screening for patients. Furthermore, they offer the opportunity for repeated tests that could potentially allow to monitor the development of the disease over time.

Transforming liquid biopsies into a standard diagnostic routine also presents many challenges (see e.g. [34]). For example, the level of biomarker in blood samples from cancer patients might be extremely low, and can thus be detected only by technologies with high sensitivity and sensibility. Furthermore, even if the biomarker is detected and analyzed successfully, the correlation between its properties and cancer features such as the tumor burden, localization, or stage need to be quantified precisely.

When a cell dies it can release small fragments of DNA that start to circulate independently in the bloodstream and are thus called cell-free DNA (cfDNA). If the cell is malignant, the shed fragments will carry some of the tumor mutations and are then referred to as circulating tumor DNA. Circulating tumor DNA (ctDNA) is one of the most commonly employed cancer biomarkers in studies on liquid biopsies, together with circulating tumor cells (CTCs) and exosomes [219, 153, 141]. In this chapter we introduce a fully stochastic model of biomarker shedding and apply it in particular to ctDNA. Intuitively, if a patient has a bigger tumor, a higher number of cancerous cells undergo apoptosis per unity of time and so we would expect increased levels of ctDNA in the blood. The main aim of our model is to provide a mechanistic description of the dynamics of cancer growth and ctDNA shedding that helps to quantify the correlation between tumor burden and ctDNA concentration.

The rest of this chapter is organized as follows. In Section 4.2 we introduce our stochastic model for tumor growth ctDNA shedding. This first setup is based on the assumption that ctDNA shedding occurs exclusively at cell apoptosis. Throughout the chapter we apply our model to lung cancer, and estimates for our model parameters in this case are presented here.
Next, in Section 4.3 we derive the probability distribution of the level of ctDNA in the bloodstream when the primary tumor is made of M cells. Even if the primary tumor is detected very early, it is still made of tens of thousands of cells,

and we thus compute these distributions in the asymptotic limit for large M.
In Section 4.4 we then relax some of our assumptions and present a more general framework that can be applied to different contexts. Specifically, we discuss the scenario where a biomarker is shed not only at cell apoptosis but also at proliferation or per unit of time. Furthermore, for other applications asymptotic results for large times, or simply the exact probability distributions of the processes involved, might be of interest and we thus derive them here.
Finally, Section 4.5 presents conclusive remarks.

4.2 Model setup

Our setup is illustrated by Figure 4.1, that can thus be used as a reference. We model the dynamics of tumor growth and biomarker shedding through a continuous-time multi-type branching process [20, 112, 23, 11, 53]. The primary tumor grows stochastically from a single malignant cell at time $t = 0$. The tumor size at time t, denoted A_t and expressed in number of cells, is modelled by a supercritical branching birth-death process with net growth rate $\lambda_A = \alpha_A - \beta_A > 0$, where α_A and β_A denote the birth and death rate per day, respectively. Adapting the notation and results in Section 1.2.1, we see that this process has a probability $q_A = \beta_A/\alpha_A$ to get extinct. Since we are only interested in primary tumors that do not go extinct, we condition A_t on eventual survival, i.e. on the event $\Omega_\infty^A := \{A_t > 0 \text{ for all } t\}$. Each cancer cell releases biomarker molecules into the bloodstream at rate ν_A per day. In this setup we assume that the biomarker is exclusively shed by cells undergoing apoptosis and hence the shedding rate is given by $\nu_A = \beta_A \cdot \mu_A$ where μ_A denotes the shedding probability per cell death.

Normal cells can also shed the biomarker into the bloodstream, as cells in benign tumors and expanded subclones often acquire the same cancer-associated mutations as cancer cells [131, 129, 186, 214]. We define B_t as the size of the benign population of cells alive at time t and shedding the same biomarker as cancer cells. We assume that healthy cells divide and die at the same rate β_B, and that the process B_t stays constant over time ($B_t = B_0$ for all t). Hence, each healthy cell sheds biomarker molecules into the bloodstream at rate $\nu_B = \beta_B \cdot \mu_B$ per day, where μ_B denotes the shedding probability per cell death. Hereafter we will assume that the shedding probabilities of cancer and normal cells are equal, i.e. $\mu_A = \mu_B = \mu$. Nevertheless, we expect the shedding rate ν_B to be lower than ν_A because normal cells often replicate at a lower rate than cancer cells [163].

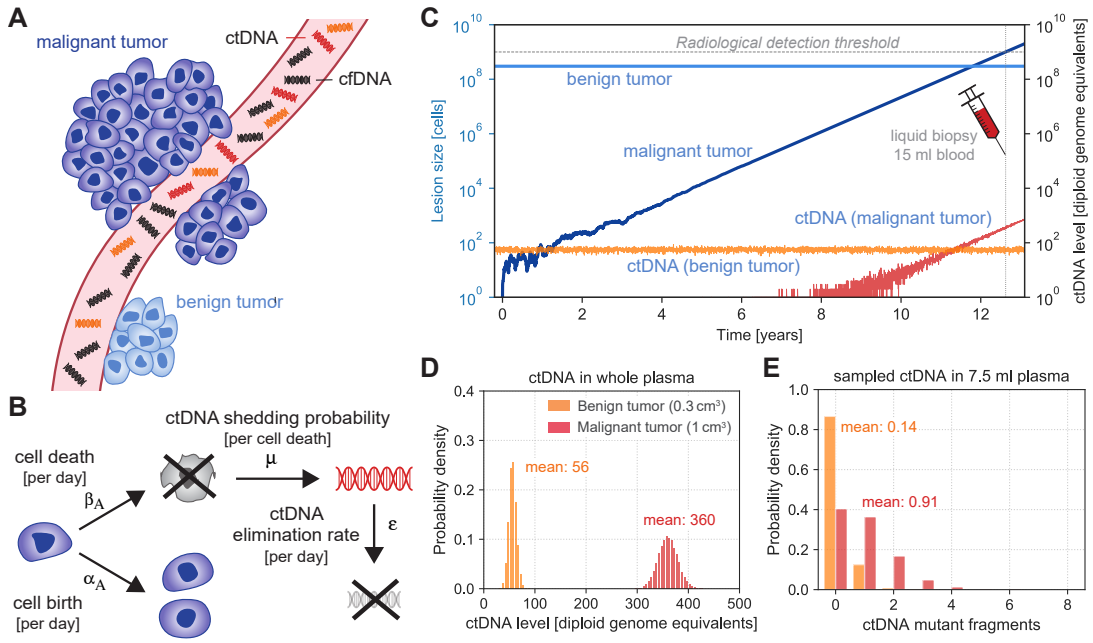

Figure 4.1: ***Evolutionary dynamics of ctDNA shed by benign and malignant tumors.*** ***A*** | *Normal cells release cell-free DNA (cfDNA) into the blood stream. Cells from benign and malignant tumors shed cell-free circulating tumor DNA (ctDNA) into the blood stream.* ***B*** | *Tumor cells divide with birth rate α_A and die with death rate β_A per day. During cell apoptosis, cells release ctDNA into the vasculature with probability μ. ctDNA is eliminated from the blood stream with rate ε per day according to the half-life time of ctDNA $t_{1/2}$.* ***C*** | *A benign lesion replicates with the same birth and death rates per day, $\alpha_B = \beta_B$. A primary tumor starts to grow at time zero with net growth rate $\lambda_A = \alpha_A - \beta_A$. Both lesions shed ctDNA into the blood stream according to the product of their cell death rate and the shedding probability μ.* ***D*** | *Distribution of ctDNA whole genome equivalents shed by the benign lesion and the primary tumor at the time when the tumor reaches a size of* 1 cm^3 *($\approx$ 1 billion cells).* ***E*** | *Probability distribution of ctDNA whole genome equivalents shed by the benign lesion and the primary tumor present in a liquid biopsy of* 15 ml *blood (*7.5 ml *plasma) at the time when the tumor reaches a size of* 1 cm^3. *The parameter values used are obtained in Section 4.2.1.*

The number of biomarker molecules in the bloodstream at time t shed by cancer and healthy cells is denoted by C_t^A and C_t^B, respectively. The total amount of biomarker at time t is thus $C_t = C_t^A + C_t^B$. Biomarker molecules are eliminated from the bloodstream at a rate ε which can be calculated from the biomarker half-life time $t_{1/2}$ as $\varepsilon = \log(2)/t_{1/2}$.

Because cancer and healthy cells shed biomarker molecules independently from each other, the processes (A_t, C_t^A) and (B_t, C_t^B) can be studied separately. The stochastic process (A_t, C_t^A) is a two-type branching process governed by the following transitions

$$\begin{aligned} A &\longrightarrow AA && \text{rate } \alpha_A\,, \\ A &\longrightarrow C && \text{rate } \beta_A \cdot \mu\,, \\ A &\longrightarrow \emptyset && \text{rate } \beta_A\,(1-\mu)\,, \\ C &\longrightarrow \emptyset && \text{rate } \varepsilon\,. \end{aligned}$$

We initialize the process at time $t = 0$ with a single cancer cell and no shed biomarkers, that is $\left(A_0, C_0^A\right) = (1, 0)$.

Since the healthy cell population B_t remains instead constant over time, C_t^B is described by a pure death process with constant immigration (see again Section 1.2.1). We assume that C_t^B is in equilibrium at time $t = 0$. Additional details about this process, including the exact form of C_0^B, will be presented in next section.

4.2.1 Parameter estimation

Before we discuss further mathematical features of our model, here we provide the parameter estimates for lung cancer that we will employ in this chapter (unless otherwise specified). These estimates are relative to the shedding of ctDNA fragments by cancerous (A_t) and healthy or pre-malignant (B_t) cell populations. The majority of parameters involved are rates, measured in cells/day. In the current setting, both the processes A_t and B_t shed the biomarker at a rate given by their cell death rate times the shedding probability ($\nu_A = \beta_A \cdot \mu$ and $\nu_B = \beta_B \cdot \mu$, respectively). Other parameters are simply measured in number of cells (initial conditions for the three processes involved) or are pure numbers (shedding probability).

For the birth rate of cancerous cells we use the estimate provided by [33, 32], $\alpha_A = 0.25$. Furthermore, [32] reports estimates for the ratio $q_A = \frac{\beta_A}{\alpha_A}$ ranging between 0.72 and 0.99. We estimate $\beta_A = 0.245$, corresponding to $q_A \approx 0.98$.

We assume that the size of the healthy cells population remains constant over time. Based on the values provided by [182] for polyp diameter, we set $B_0 = 3 \times 10^8$. Furthermore, in order to estimate how often this healthy population sheds ctDNA, we need a measure of its turnover rate. Based on the values indicated by [124], we set $\alpha_B = \beta_B = 0.1$.

As for ctDNA elimination rate ε, [199] reports values of ctDNA halftime in the bloodstream, $t_{1/2}$, ranging between 16 and 150 minutes. We employ the estimate $t_{1/2} = 30$ minutes, corresponding to $\varepsilon = \frac{\log(2)}{t_{1/2}} \approx 33.27$ ctDNA fragments per day.

Next we need to estimate the shedding probability q_A, which in turn determines the shedding rates ν_A and ν_B. We assume that the shedding probability μ is independent of cancer type. Bidard et al. [25] reported a count of 1 copy of ctDNA shed per cm^3 of metastatic uveal melanoma per ml of plasma. Similarly, Parkinson et al. [150] estimated 6 ctDNA copies in high-grade serous ovarian carcinoma and Abbosh et al. [2] estimated 0.12 ctDNA copies shed per per cm^3 of non-small cell lung cancer (NSCLC) per ml of plasma (increase of 0.008% in VAF per cm^3, assuming 3000 whole genome equivalents per ml of plasma in healthy subjects) [199, 63]. Since the NSCLC cohort was mostly composed of stage I and stage II cancers while the cancers in the other cohorts were at stage III or IV, we used the reported estimate in the NSCLC cohort. The value inferred from a cohort of early stage cancers has indeed lower chances to be affected by undetected metastases, that as discussed in Chapter 3 are more likely present at later primary stages and can greatly increase the total tumor burden. Assuming an average patient has roughly 6 liters of blood and that half of it is plasma, the conservative estimate of 0.12 ctDNA whole genome equivalents per ml of plasma shed per cm^3 of tumor volume leads to 360 ($= 0.12 \cdot 3000$) copies of ctDNA in the whole plasma. As 1 cm^3 of tumor volume contains 10^9 cells, we can then setup the equilibrium equation

$$\nu_{\text{NSCLC}} \cdot 10^9 = 360 \cdot \varepsilon$$

which leads to a shedding rate of $\nu_{\text{NSCLC}} \approx 1.198 \times 10^{-5}$ per cell per day. Because ctDNA shedding largely occurs during cell apoptosis [84] and lung cancer

cells have been measured to die every ≈ 7 days [163], we find that the shedding probability is $\mu = \nu_{\text{NSCLC}}/\beta_{\text{NSCLC}} \approx 8.38 \times 10^{-5}$ per cell death.
To calculate the shedding rate of cells of other cancer types or tissues, we can now simply rescale the shedding probability by the corresponding cell death rate of these cells. For example, we estimate the shedding rate of benign cells by $\nu_B = \beta_B \cdot \mu$.

To conclude this section let us stress that the parameters inferred above must be interpreted as typical values for the corresponding biological quantities. While the results presented in the following are based on these estimates, in prospect we aim to refine such inference - as more data become available - and possibly to derive a probability distribution for the parameter values. Finally, for a discussion of the sensitivity of our model to these values we refer to the later Section 4.3.4, and in particular to Figure 4.2.

4.3 ctDNA level distributions

The processes A_t, B_t and C_t are indexed by the time parameter t, where $t = 0$ at the primary tumor onset. In practical situations this time is impossible to measure, and it is thus convenient to express the previous processes in terms of other indexing parameters. In our case a suitable choice is to use the first passage time

$$T_M := \inf\{t > 0 : A_t = M\}\,.$$

Indeed, while the behaviour of the processes A_t and B_t at T_M is obvious, $C^A_{T_M}$ and $C^B_{T_M}$ express ctDNA levels in terms of an observable quantity, i.e. the primary tumor size. Moreover, we are especially interested in our model predictions for relatively large primary tumor sizes. When M is large and the shedding rate is small, we are able to apply asymptotic results for the distributions of the processes $C^A_{T_M}$ and $C^B_{T_M}$ that greatly simplify computations and help getting insight into the main evolutionary features of the model. In the following we thus focus on these results first. Later we will generalize our model to include multiple biomarker shedding dynamics and derive the probability distribution of the processes A_t, B_t and C_t indexed by time.

4.3.1 ctDNA shed by cancerous cells

The asymptotic behaviour of $C^A_{T_M}$ in the small shedding rate - large primary tumor size limit can be derived from referring to [37]. Adapting the results presented there, we first observe that ctDNA fragments are shed by cancer cells at a stochastic rate which is proportional to the primary tumor size, i.e. as a Cox process with intensity $(\nu_A A_t)_{t\geq 0}$ (see Section 1.2.3). For large times the branching birth-death process A_t satisfies the classic result given by equation (1.11)

$$\lim_{t\to\infty} A_t = W_A e^{\lambda_A t}\,. \tag{4.1}$$

Here, we recall that W_A is a non-negative random variable such that $W_A = 0$ if and only if the process A_t goes extinct. Using these properties and the definition of T_M, one can prove that in the small ν_A - large M limit and conditioning on $\Omega^A_\infty = \{W > 0\}$ the number of ctDNA fragments shed by A_t up to time T_M converges in distribution to a Poisson random variable with rate $\frac{\nu_A M}{\lambda_A}$. Furthermore, as the age and the lifespan of a randomly selected fragment are $\mathrm{Exp}(\lambda_A)$ and $\mathrm{Exp}(\varepsilon)$ random variables respectively, the probability that any of the shed ones is still present in the bloodstream at time T_M is $\frac{\lambda_A}{\varepsilon+\lambda_A}$. Hence, due to the thinning property of Poisson processes (see equation (1.22)) we find that for large primary tumor size M and small shedding rate ν_A

$$C^A_{T_M} \sim \mathrm{Pois}\left(\frac{\nu_A \cdot M}{\varepsilon + \lambda_A}\right)\,. \tag{4.2}$$

4.3.2 ctDNA shed by healthy cells

As mentioned in the previous section, if B_t is constant over time the process C^B_t is a branching pure death process with immigration, with death rate ε and immigration rate $B_0\nu_B = B_0\beta_B\mu$. The probability generating function for such a process thus follows from equation (1.15) by taking the limit for the birth rate that tends to zero

$$C^B(y,t) = \left(1 + (y-1)e^{-\varepsilon t}\right)^{C^B_0} e^{\frac{\nu_B}{\varepsilon}B_0(y-1)(1-e^{-\varepsilon t})}\,. \tag{4.3}$$

Now, conditioning on A_t survival we have that $\lim_{M\to\infty} T_M = \infty$ almost surely. Hence, as the process (B_t, C^B_t) is independent of A_t, the large M limit of $C^B_{T_M}$ corresponds to the large time limit of C^B_t. When t is large, equation (4.3) converges to

$$\lim_{t\to\infty} C^B(y,t) = e^{\frac{\nu_B}{\varepsilon}B_0(y-1)}\,,$$

independently of the initial number of ctDNA fragments. The right hand side in the last equation is the probability generating function of a Poisson random variable with mean $\frac{\nu_B}{\varepsilon}B_0$, and so for large M we have

$$C^B_{T_M} \sim \text{Pois}\left(\frac{\nu_B}{\varepsilon}B_0\right) . \tag{4.4}$$

As we assume that C^B_t is originally at equilibrium, we set $C^B_0 = \frac{\nu_B}{\varepsilon}B_0$.

4.3.3 Total ctDNA levels and sampling

By joining the previous results we find the asymptotic distribution for the total amount of ctDNA present in the bloodstream when the primary tumor is made of a large number M of cells. In particular we get that, conditioned on A_t survival,

$$C_{T_M} = C^A_{T_M} + C^B_{T_M} \xrightarrow{d} \text{Pois}\left(\frac{M\,\beta_A\,\mu}{\varepsilon+\lambda_A} + \frac{B_0\,\beta_B\,\mu}{\varepsilon}\right) . \tag{4.5}$$

Now, suppose that a blood sample of volume V_s is taken from a patient with a total blood volume equal to V_{tot}. Assuming that the extant ctDNA fragments are well mixed in the bloodstream, we approximate that each of them is present in the sample with probability $p = V_s/V_{tot}$. Hence, by denoting X_t the number of fragments in the sample at time t, we can apply again the thinning property of Poisson processes and find

$$X_{T_M} \sim \text{Pois}\left(p\left[\frac{M\,\beta_A\,\mu}{\varepsilon+\lambda_A} + \frac{B_0\,\beta_B\,\mu}{\varepsilon}\right]\right) . \tag{4.6}$$

4.3.4 ctDNA shedding dynamics for lung cancer

We now discuss some quantitative and qualitative features of our model by applying the estimates for lung cancer to the results we derived in the previous section. Using such estimates (obtained in Section 4.2.1), Figure 4.2 highlights the dependencies of ctDNA levels on different parameters.

Our model predicts that smaller tumors lead to lower levels of ctDNA in the bloodstream. We estimate that tumors with 300 million cells and a billion cells (≈ 0.3 cm^3 and ≈ 1 cm^3) correspond to around 200 and 650 ctDNA fragments present, respectively (Figure 4.2A). On the other hand, among two cancers of the same size, the more aggressive one also leads to lower levels of ctDNA (Figure 4.2B) because fewer cell deaths decrease the amount of released biomarker. For the same reason, high cell turnover rates correspond instead to a higher level

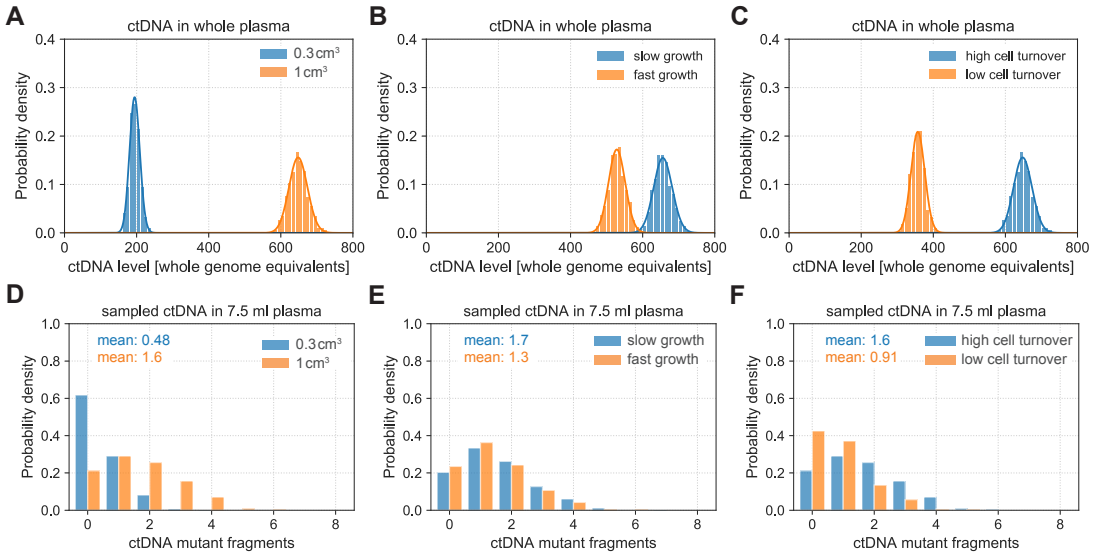

Figure 4.2: ***Tumor growth rate and cell turnover strongly affect the amount of ctDNA.*** *Top panels (**A-C**) depict the number of ctDNA fragments present in 3 liters of plasma at a given tumor size. Bars illustrate the distribution of ctDNA levels based on 10^4 simulations. Full lines plot instead the asymptotic results derived from equation (4.5), which perfectly agree with simulation results. Bottom panels (**D-F**) show the simulated distribution of the number of ctDNA fragments present in 7.5 ml of plasma, corresponding to a standard 15 ml liquid biopsy.*

of ctDNA at a given tumor size (1 billion cells in Figure 4.2C) compared to lower cell turnover rates.

These features reflect on the distributions of ctDNA levels in plasma samples. Panels D-F in Figure 4.2 show the probability densities derived from equation (4.6) for a blood sample of 15 ml taken at the time when the tumors reach the given sizes. From these plots we observe that the tumor size has the highest impact on the expected number of ctDNA fragments present in the sample. In fact, for all tumors made of a billion cells this expected value is higher than or close to 1, and it is considerably lower only for the simulated tumor with volume ≈ 0.3 cm^3.

These latter considerations are particularly interesting for early detection. In fact, even if highly sensible technologies for the analysis of blood samples are available, they will not be able to detect reliably more than a few ctDNA fragments in a liquid biopsy. In Figure 4.3 we thus show the probabilities of at least one, two and three ctDNA fragments present in a 7.5 ml plasma sample, again as functions of different parameters.

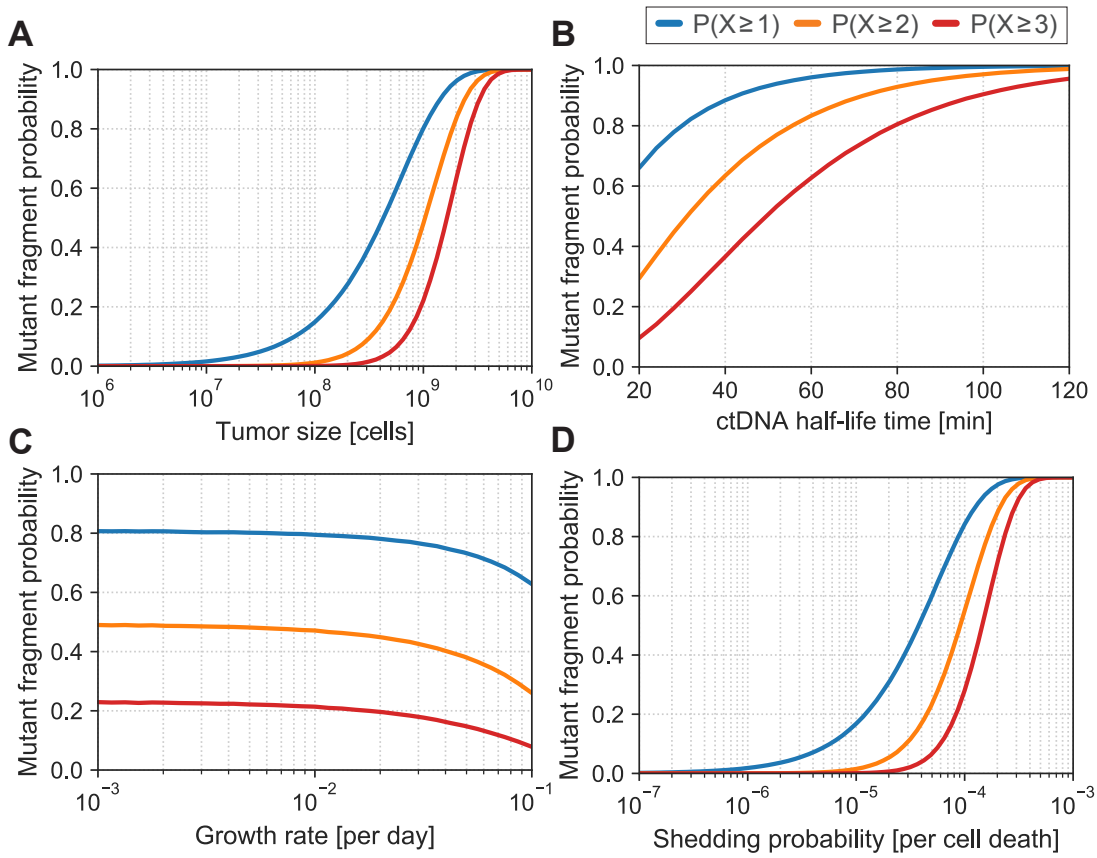

Figure 4.3: ***Detection probability of ctDNA fragments in a 15 ml liquid biopsy for a given tumor depends on many parameters.*** *Full lines denote the probability to find at least 1 (blue), 2 (orange), or 3 (red) mutant ctDNA fragments in a liquid biopsy of* 15 ml *blood (≈* 7.5 ml *plasma).*

The plots depicted in Figure 4.3A confirm that tumors with fewer than 10^9 cells rarely shed a detectable amount of ctDNA fragments. We then observe that, at a given tumor size, higher ctDNA half-life time corresponds to a lower elimination rate from the bloodstream and thus increases the number of mutant ctDNA fragments (Figure 4.3B). Similarly, the ctDNA detection probabilities quickly increase with the shedding probability at cell death (Figure 4.3D). On the other hand, Figure 4.3C confirms that slowly growing tumors likely yield a higher number of ctDNA fragment in a plasma sample than fast growing ones.

4.4 General Framework

The results we have derived so far to describe ctDNA shedding by cancerous and healthy cells can be extended in several ways. Firstly, so far we have focussed on the case where biomarker shedding happens only at cell apoptosis, but our model

can be generalized to include shedding also at cell necrosis (per unity of time) and proliferation. Furthermore, we presented size distributions in the asymptotic limit of large primary tumor sizes - small shedding rates, and similar distributions can be obtained also in the asymptotic limit for large times - small shedding rates. Finally, even taking into account multiple shedding dynamics it is possible to compute exact probability distributions for the processes A_t, B_t and C_t at any given time t. In the following we will discuss in greater details these three generalizations.

In order to include in our model the possibility of biomarker shedding by cancerous cells also at necrosis and proliferations, we consider the following set of transitions for the process (A_t, C_t)

$$\begin{array}{lll} A \longrightarrow AAC & \text{rate } \alpha_A\,\mu_\alpha\,, & \text{rate } \nu_2\,, \\ A \longrightarrow AA & \text{rate } \alpha_A\,(1-\mu_\alpha)\,, & \text{rate } \alpha_A - \nu_2\,, \\ A \longrightarrow AC & \text{rate } \nu_1\,, & \text{rate } \nu_1\,, \\ A \longrightarrow C & \text{rate } \beta_A\,\mu_\beta\,, & \text{rate } \nu_0\,, \\ A \longrightarrow \emptyset & \text{rate } \beta_A\,(1-\mu_\beta)\,, & \text{rate } \beta_A - \nu_0\,, \\ C \longrightarrow \emptyset & \text{rate } \varepsilon\,, & \text{rate } \varepsilon\,. \end{array} \tag{4.7}$$

Here, μ_β and μ_α represent the shedding probabilities at apoptosis and proliferation, respectively. Similarly, $\nu_0 = \beta_A\,\mu_\beta$, ν_1 and $\nu_2 = \alpha_A\,\mu_\alpha$ denote the shedding rates at cancerous cells apoptosis, necrosis and reproduction. The total shedding rate is then defined by $\nu_A = \nu_0 + \nu_1 + \nu_2$.

4.4.1 Asymptotic results with multiple shedding dynamics

With the updated definition of the total shedding rate ν_A, we still have that biomarker molecules are shed by cancer cells as a Cox process with intensity $(\nu_A A_t)_{t\geq 0}$. Hence, the derivation of $C^A_{T_M}$ distribution follows exactly as before. In particular, we find again that for a large primary tumor size M and a small total shedding rate ν_A, the number of biomarker molecules present in the bloodstream when $A_t = M$ is approximately a Poisson random variable with mean $\frac{\nu_A\,M}{\varepsilon+\lambda_A}$.

Asymptotic results for large times

A similar derivation provides C^A_t size distribution in the asymptotic limit for large time and small total shedding rate. Following again the results presented

in [37], we see that under this limit and conditioning on A_t survival the total number of biomarker molecules shed up to time t converges in finite dimensional distributions to a Poisson random variable with mean $\frac{W_A \nu_A e^{\lambda_A t}}{\lambda_A}$, where W_A is the same as in equation (4.1). In particular, we recall that, conditioned on A_t non-extinction, W_A follows an Exp $\left(\frac{\lambda_A}{\alpha_A}\right)$ distribution (see equation (1.13)). The number of shed biomarker molecules still present in the bloodstream at time t is thus a compound Poisson process with probability generating function

$$\mathbb{E}\left[z^{C_t^A}\right] = \mathbb{E}\left[\exp\left(\frac{W_A \nu_A e^{\lambda_A t}}{\lambda_A}\left(r(z)-1\right)\right)\right].$$

Here $r(z)$ denotes the probability generating function of the process initiated by a randomly selected molecule, but as the biomarker cannot reproduce this is a pure death branching process whose size can only be 1 or 0. The probability generating function of such a process is $r(z) = \frac{\varepsilon + \lambda_A z}{\varepsilon + \lambda_A}$ (see equation (1.4)). Using this expression and the results for W, the probability generating function for C_t^A becomes

$$\mathbb{E}\left[z^{C_t^A}\right] = \left(1 - \frac{\alpha_A \nu_A e^{\lambda_A t}(z-1)}{\lambda_A(\varepsilon + \lambda_A)}\right)^{-1}.$$

Notice that the same steps can be repeated to derive more formally our asymptotic results for large primary tumor sizes. We recognize the last expression as the probability generating function of a geometric random variable, and so for large times and small total shedding rates, conditioned on A_t survival, we find

$$C_t^A \sim \text{Geometric}\left(\frac{\lambda_A(\varepsilon + \lambda_A)}{\alpha_A \nu_A e^{\lambda_A t} + \lambda_A(\varepsilon + \lambda_A)}\right). \tag{4.8}$$

Expected values and variances for the processes A_t, C_t^A and C_t^B are summarized in Appendix C.1.

4.4.2 Exact results

The asymptotic results previously derived provide a relatively simple toolbox to quantify the correlation between ctDNA levels and tumor burden. In principle, however, our mathematical framework can be used for other biomarkers - or entirely different applications - where the model parameters do not fall in the asymptotic regimes considered so far. For this reason we now show how to derive the exact distributions of the processes A_t, B_t and C_t, for any given time t and combinations of shedding rates.

Probability generating functions

In order to compute the exact distribution of A_t, B_t and C_t we first derive the joint probability generating functions of the processes (A_t, C_t^A) and (B_t, C_t^B). The marginal generating functions then follow straightforwardly, and provide the probability distributions of the single processes by simple analytical or numerical inversion.

In order to simplify the notation, where no ambiguity is caused we will hereafter omit the subscripts A and B for the model parameters.

Process (A_t, C_t^A)**.** Our derivation for the process (A_t, C_t^A) is adapted from that in [13, 14], which first obtained similar results for a two-type branching process where the second type has the ability to reproduce but shedding (or mutations) can happen only at wild-type cell death. As before, the process is completely defined by the set of transitions (4.7). For any given initial condition let us denote $P^*_{m,n}(t) = P((A_t, C_t^A) = (m,n) \mid (A_0, C_0^A) = *)$. The corresponding probability generating function is then defined as

$$\boldsymbol{\mathcal{P}}^*(x,y,t) = \sum_{m,n\geq 0} x^m y^n P^*_{m,n}(t)\,.$$

Now, the backward Kolmogorov equations for our model read

$$\begin{aligned}
\frac{dP^{(1,0)}_{m,n}}{dt} &= (\alpha-\nu_2)P^{(2,0)}_{m,n} + (\beta-\nu_0)\delta_{m,0}\delta_{n,0} \\
&\quad + \nu_0 P^{(0,1)}_{m,n} + \nu_1 P^{(1,1)}_{m,n} + \nu_2 P^{(2,1)}_{m,n} - (\alpha+\beta+\nu_1)P^{(1,0)}_{m,n}\,, \qquad (4.9)\\
\frac{dP^{(0,1)}_{m,n}}{dt} &= \varepsilon\,\delta_{m,0}\delta_{n,0} - \varepsilon P^{(0,1)}_{m,n}\,.
\end{aligned}$$

Multiplying both sides of equation (4.9) by $x^m y^n$ and summing over all non-negative m, n we find

$$\begin{aligned}
\partial_t \boldsymbol{\mathcal{P}}^{(1,0)} &= (\alpha-\nu_2)\boldsymbol{\mathcal{P}}^{(2,0)} + (\beta-\nu_0) \\
&\quad + \nu_0 \boldsymbol{\mathcal{P}}^{(0,1)} + \nu_1 \boldsymbol{\mathcal{P}}^{(1,1)} + \nu_2 \boldsymbol{\mathcal{P}}^{(2,1)} - (\alpha+\beta+\nu_1)\boldsymbol{\mathcal{P}}^{(1,0)}\,,\\
\partial_t \boldsymbol{\mathcal{P}}^{(0,1)} &= \varepsilon - \varepsilon \boldsymbol{\mathcal{P}}^{(0,1)}\,.
\end{aligned}$$

Next, we notice that

$$\boldsymbol{\mathcal{P}}^{(2,0)}(x,y,t) = \left[\boldsymbol{\mathcal{P}}^{(1,0)}(x,y,t)\right]^2\,,$$

because of the independence of the progenies of the two initial cells. By applying this property and introducing the notation

$$\mathcal{A}(x,y,t) = \mathcal{P}^{(1,0)}(x,y,t), \quad \mathcal{C}(x,y,t) = \mathcal{P}^{(0,1)}(x,y,t)\,,$$

we reduce to the following system of first order differential equations

$$\partial_t \mathcal{A} = (\alpha - \nu_2)\mathcal{A}^2 + (\beta - \nu_0) + \nu_0 \mathcal{C} + \nu_1 \mathcal{A}\mathcal{C} + \nu_2 \mathcal{A}^2 \mathcal{C} - (\alpha + \beta + \nu_1)\mathcal{A}, \tag{4.10}$$

$$\partial_t \mathcal{C} = \varepsilon - \varepsilon \mathcal{C}\,, \tag{4.11}$$

with initial conditions

$$\mathcal{A}(x,y,0) = x\,, \tag{4.12}$$

$$\mathcal{C}(x,y,0) = y\,. \tag{4.13}$$

Equation (4.11), subject to the initial condition (4.13), has solution

$$\mathcal{C}(x,y,t) = 1 + (y-1)e^{-\varepsilon t}\,. \tag{4.14}$$

Plugging this back into equation (4.10) we get

$$\partial_t \mathcal{A} = \left[\alpha + \nu_2(y-1)e^{-\varepsilon t}\right]\mathcal{A}^2 + \left[\nu_1(y-1)e^{-\varepsilon t} - \alpha - \beta\right]\mathcal{A} + \left[\beta + \nu_0(y-1)e^{-\varepsilon t}\right].$$

We first apply the change of variables $s = e^{-\varepsilon t}$ and find

$$\partial_s \mathcal{A} = -\frac{[\alpha + \nu_2(y-1)s]}{\varepsilon s}\mathcal{A}^2 - \frac{[\nu_1(y-1)s - \alpha - \beta]}{\varepsilon s}\mathcal{A} - \frac{[\beta + \nu_0(y-1)s]}{\varepsilon s}\,.$$

To ease the notation we rewrite the equation above as

$$\partial_s \mathcal{A} = \frac{a_2 + b_2 s}{s}\mathcal{A}^2 + \frac{a_1 + b_1 s}{s}\mathcal{A} + \frac{a_0 + b_0 s}{s}\,, \tag{4.15}$$

where

$$a_2 = -\frac{\alpha}{\varepsilon}, \quad a_1 = \frac{\alpha+\beta}{\varepsilon}, \quad a_0 = -\frac{\beta}{\varepsilon}\,,$$
$$b_2 = -\frac{\nu_2(y-1)}{\varepsilon}, \quad b_1 = -\frac{\nu_1(y-1)}{\varepsilon}, \quad b_0 = -\frac{\nu_0(y-1)}{\varepsilon}\,.$$

Notice that we are interested in the probability generating function $\mathcal{A}(x,y,t)$ for $t \geq 0$, that corresponds to $0 < s \leq 1$. Equation (4.15) is a Riccati equation, which can be reduced to a second order ODE. In order to do so, we define

$$X(x,y,s) = \frac{a_2 + b_2 s}{s} \mathcal{A}(x,y,s)\,,$$

which yields

$$\partial_s X = X^2 + \frac{(a_2 + b_2 s)(a_1 + b_1 s) - a_2}{s(a_2 + b_2 s)} X + \frac{(a_2 + b_2 s)(a_0 + b_0 s)}{s^2}\,. \tag{4.16}$$

Next, we set

$$X \equiv -\frac{\partial_s Y}{Y} \quad \Longrightarrow \quad \partial_s X = \frac{(\partial_s Y)^2}{Y^2} - \frac{\partial_s^2 Y}{Y}\,,$$

so that equation (4.16) transforms into

$$\partial_s^2 Y - \frac{(a_2 + b_2 s)(a_1 + b_1 s) - a_2}{s(a_2 + b_2 s)} \partial_s Y + \frac{(a_2 + b_2 s)(a_0 + b_0 s)}{s^2} Y = 0\,. \tag{4.17}$$

Equation (4.17) is a second order linear differential equation with rational coefficients. It has three singular points - $s = 0$ and $s = -\frac{a_2}{b_2}$ being regular and $s = \infty$ being irregular with rank 1. Hence, equation (4.17) is a single confluent Heun equation [164]. To bring it into standard form we first move the non-zero regular singularity to 1 through the change of variable $z = -\frac{b_2}{a_2} s$. This leads to

$$\partial_z^2 Y + \frac{(z-1)\left(\frac{a_2 b_1}{b_2} z - a_1\right) - 1}{z(z-1)}\, \partial_z Y + \frac{a_2 (z-1)\left(\frac{a_2 b_0}{b_2} z - a_0\right)}{z^2}\, Y = 0\,. \tag{4.18}$$

We now look for a solution to equation (4.18) of the form

$$Y(z) = z^m e^{nz} f(z)\,, \tag{4.19}$$

which implies

$$\begin{aligned} \partial_z Y &= z^{m-1} e^{nz} \left[(m + nz) f + z f'\right]\,, \\ \partial_z^2 Y &= z^{m-2} e^{nz} \left\{ \left[(m+nz)^2 - m\right] f + 2z(m+nz) f' + z^2 f'' \right\}\,. \end{aligned}$$

Substituting these expressions into equation (4.18) and rearranging, we reduce to the following differential equation for $f(z)$

$$f'' + \frac{P(z)}{z(z-1)} f' + \frac{Q(z)}{z^2(z-1)} f = 0\,, \tag{4.20}$$

where $P(z)$ and $Q(z)$ are two polynomials of second and third degree in z, respectively. However, it is easily checked that by taking

$$m = -a_2, \quad n = -\frac{a_2}{2b_2}\left[b_1 + \sqrt{b_1^2 - 4b_0 b_2}\right], \tag{4.21}$$

the first and last coefficients of the polynomial $Q(z)$ become zero. Hence, for these values equation (4.20) can be written as

$$f'' + \left(\frac{\gamma}{z} + \frac{\delta}{z-1} + \eta\right) f' + \frac{\omega z + \rho}{z(z-1)} f = 0\,. \tag{4.22}$$

Here, the parameters γ, δ and η follow from the decomposition in partial fractions of $\frac{P(z)}{z(z-1)}$ while ω and ρ correspond to the coefficients of second and first degree terms in $Q(z)$, respectively - when m and n are set as in equation (4.21). Explicit expressions for all these parameters in terms of the original rates are given by

$$\begin{aligned}
m &= \frac{\alpha}{\varepsilon}, \qquad n = \frac{\alpha}{2\nu_2\varepsilon}\left(\nu_1 - \sqrt{\nu_1^2 - 4\nu_0\nu_2}\right), \\
\gamma &= \frac{\alpha - \beta}{\varepsilon} + 1\,, \qquad q = -1\,, \qquad \eta = \frac{\alpha}{\nu_2\varepsilon}\sqrt{\nu_1^2 - 4\nu_0\nu_2}\,, \\
\omega &= -\frac{\alpha}{2\nu_2\varepsilon^2}\left[(\alpha+\beta)\nu_1 - (\alpha-\beta)\sqrt{\nu_1^2 - 4\nu_0\nu_2} + 2(\alpha\nu_0 + \beta\nu_2)\right], \\
\rho &= \frac{\alpha}{2\nu_2\varepsilon^2}\left\{(\alpha+\beta-\varepsilon)\nu_1 - (\alpha-\beta+\varepsilon)\sqrt{\nu_1^2 - 4\nu_0\nu_2} + 2[\alpha\nu_0 + (\beta-\varepsilon)\nu_2]\right\}.
\end{aligned} \tag{4.23}$$

Now, equation (4.22) is the standard form of the single confluent Heun equation. Solutions to such a second order differential equation are called confluent Heun functions and depend on six arguments - the five parameters of the equation and the independent variable z. A standardized package for numerical and symbolical computations involving Heun functions is currently provided only by Maple [136], which includes in particular the procedures `HeunC` and `HeunCPrime` for the evaluation of confluent Heun functions and their z derivative, respectively. Consistently with these implementations the general solution of equation (4.22) can be written as

$$f(z) = h_1(z) + D z^{1-\gamma} h_2(z)\,, \tag{4.24}$$

where

$$h_1(z) = \texttt{HeunC}\left(\eta, 1-\gamma, \rho-1, \omega - \frac{\eta(\gamma+\rho)}{2}, \rho + \frac{1-\gamma(\rho-\eta)}{2}, z\right),$$
$$h_2(z) = \texttt{HeunC}\left(\eta, \gamma-1, \rho-1, \omega - \frac{\eta(\gamma+\rho)}{2}, \rho + \frac{1-\gamma(\rho-\eta)}{2}, z\right),$$

are the confluent Heun functions uniquely determined by the following initial conditions

$$\begin{aligned} h_1(0) &= 1, \qquad h_1'(0) = \frac{\rho}{\gamma}, \\ h_2(0) &= 1, \qquad h_2'(0) = \frac{(\gamma-1)(\rho-\eta)-\rho}{\gamma-2}. \end{aligned} \tag{4.25}$$

Here h_i' denotes the z derivative of the confluent Heun functions h_i. As mentioned before these are implemented by the Maple procedure `HeunCPrime` and uniquely determined by an additional condition on $h_i''(0)$, too cumbersome to be reported here. Also notice that the expression in equation (4.24) contains only one integrating constant, D, as our derivation spawns from a first order differential equation. By plugging such expression back into equation (4.19), the solution to equation (4.18) becomes

$$Y(z) = z^m e^{nz} f(z) = z^m e^{nz} h_1(z) + D z^{1+m-\gamma} e^{nz} h_2(z).$$

Next, we denote the derivative of $f(z)$ as

$$g(z) = f'(z) = h_1'(z) + D z^{-\gamma}[(1-\gamma)h_2(z) + z h_2'(z)]. \tag{4.26}$$

Hence, recalling that $z = -\frac{b_2}{a_2}s = ks$ and that $X(x,y,s) = -\frac{\partial_s Y}{Y}$, we find

$$X(x,y,s) = -\frac{(m+nks)f(ks) + ksg(ks)}{sf(ks)}. \tag{4.27}$$

Notice that in terms of the original parameters of our model the coefficient k is given by $k = \mu_\alpha(y-1)$ and thus depends on y. We can now apply the initial condition for $X(x,y,s)$ to find the value of the constant D. Since $s = e^{-\varepsilon t}$ and $X(x,y,s) = \frac{a_2+b_2 s}{s}\mathcal{A}(x,y,s)$, the initial condition $\mathcal{A}(x,y,t=0) = x$ translates to $X(x,y,s=1) = (a_2+b_2)x = a_2(1-k)x$. Therefore equation (4.27) at $s=1$ implies

$$a_2(1-k)x = -\frac{(m+nk)f(k) + kg(k)}{f(k)}.$$

Substituting the expressions for f and its derivative (equations (4.24) and (4.26), respectively) and solving for D we find

$$D = -\frac{(m + nk + a_2x - ka_2x)h_1(k) + kh_1'(k)}{k^{1-\gamma}[(m + nk + a_2x - ka_2x + 1 - \gamma)h_2(k) + kh_2'(k)]}\,.$$

Hence, plugging this value back into equation (4.27) we find an expression for X in terms of the functions $h_i(s)$ and $h_i'(s)$. Multiplying such expression by $\frac{s}{a_2(1-ks)}$ and substituting $s = e^{-\varepsilon t}$ we eventually find

$$\mathcal{A}(x, y, t) = \frac{e^{\lambda t}K_1(x, y)\phi_2(y, t) - K_2(x, y)\phi_1(y, t)}{e^{\lambda t}K_1(x, y)\psi_1(y, t) - K_2(x, y)\psi_2(y, t)}\,, \tag{4.28}$$

where

$$\begin{aligned}
K_1(x, y) &= \left\{\frac{\alpha - x[\alpha + \nu_2(y - 1)]}{\varepsilon} + nk\right\} h_1(k) + kh_1'(k)\,,\\
K_2(x, y) &= \left\{\frac{\beta - x[\alpha + \nu_2(y - 1)]}{\varepsilon} + nk\right\} h_2(k) + kh_2'(k)\,,\\
\phi_1(y, t) &= \left[\frac{\alpha nke^{-\varepsilon t}}{\varepsilon} + 1\right] h_1(ke^{-\varepsilon t}) + \frac{\varepsilon k}{\alpha}e^{-\varepsilon t}h_1'(ke^{-\varepsilon t})\,,\\
\phi_2(y, t) &= \left[\frac{\alpha nke^{-\varepsilon t}}{\varepsilon} + \frac{\beta}{\alpha}\right] h_2(ke^{-\varepsilon t}) + \frac{\varepsilon k}{\alpha}e^{-\varepsilon t}h_2'(ke^{-\varepsilon t})\,,\\
\psi_1(y, t) &= (1 - ke^{-\varepsilon t})h_1(ke^{-\varepsilon t})\,,\\
\psi_2(y, t) &= (1 - ke^{-\varepsilon t})h_2(ke^{-\varepsilon t})\,.
\end{aligned} \tag{4.29}$$

This is the joint probability generating function of the processes A_t and C_t^A, starting from one A cell at time $t = 0$.

To find the marginal generating function

$$\mathcal{A}(x, t) = \sum_{m=0}^{\infty} \mathbb{P}(A_t = m \mid (A_0, C_0^A) = (1, 0))x^m\,,$$

we take the limit for $y \to 1$ in equation (4.28). By applying the conditions in equation (4.25), we retrieve the probability generating function of a supercritical birth-death process with net growth rate $\lambda = \alpha - \beta > 0$ and extinction probability $q = \beta/\alpha < 1$, as given by equation (1.4). Similarly, the generating function

$$C^A(y, t) = \sum_{n=0}^{\infty} \mathbb{P}(C_t^A = n \mid (A_0, C_0^A) = (1, 0))y^n$$

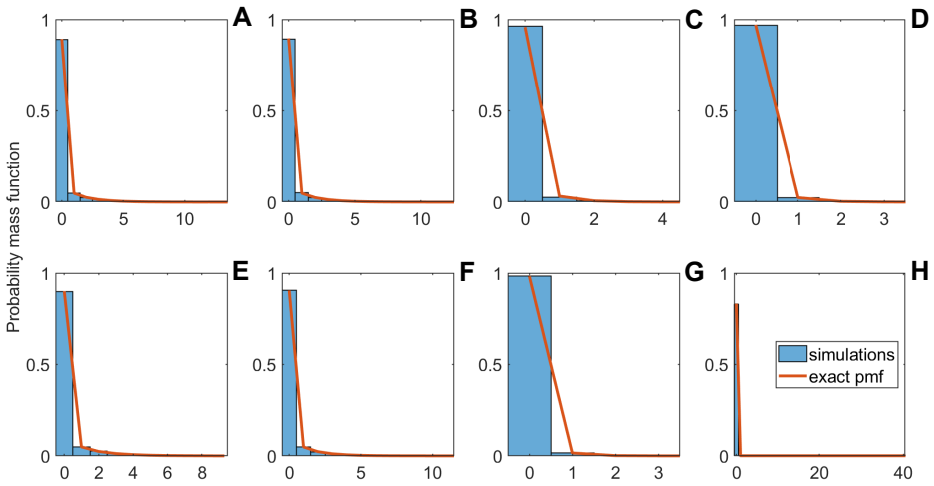

Figure 4.4: ***Comparison between simulated ctDNA levels and their exact theoretical distributions with multiple shedding dynamics.*** *Panels (**A-G**) show the level of biomarker shed by cancer cells and still circulating in the bloodstream at time t for all the possible combinations of non-zero shedding rates. Bars illustrate the distribution of the number of biomarker molecules based on 10^4 simulations. Full red lines illustrate the exact probability distribution at time t derived from equations (4.30), (C.3) and (C.6) and perfectly agree with simulation results. Panel (**H**) similarly shows the probability distribution of the primary tumor size. Parameter values: birth rate $\alpha_A = 0.051$ per cell per day and death rate $\beta_A = 0.041$ per cell per day; elimination rate $\varepsilon = 3.1 \times 10^{-5}$ per biomarker molecule per day; biomarker shedding probability per cell death $\mu_\beta = 10^{-4}$, per cell reproduction $\mu_\alpha = 1.5 \times 10^{-4}$, and shedding rate per day (during cell necrosis) $\nu_1 = 5 \times 10^{-5}$ per cell per day. All results are computed at time $t = 365$ days from the primary tumor onset.*

is derived by taking the limit for $x \to 1$ in equation (4.28). The expression for $\mathcal{C}^A$ does not simplify significantly and becomes

$$\mathcal{C}^A(y,t) = \frac{e^{\lambda t}\bar{K}_1(y)\phi_2(y,t) - \bar{K}_2(y)\phi_1(y,t)}{e^{\lambda t}\bar{K}_1(y)\psi_1(y,t) - \bar{K}_2(y)\psi_2(y,t)}\,, \tag{4.30}$$

where

$$\bar{K}_1(y) = \left\{ nk - \frac{\nu_2(y-1)}{\varepsilon} \right\} h_1(k) + k h_1'(k)\,,$$
$$\bar{K}_2(y) = \left\{ nk - \frac{\alpha - \beta + \nu_2(y-1)}{\varepsilon} \right\} h_2(k) + k h_2'(k)\,,$$

and the functions $\phi_1, \phi_2, \psi_1, \psi_2$ are defined as in equation (4.29). The function $\mathcal{C}^A(y,t)$ can be then inverted numerically (see Section 1.2.1) to find the distribution of the process C_t^A.

When one or more of the shedding rates ν_0, ν_1 and ν_2 are zero, the results derived above assume slightly different forms. For a discussion of these special cases and the corresponding probability generating functions, see Appendix C.2.1. The probability mass function of C_t^A in the general case presented here and in the special cases just mentioned is visualized by Figure 4.4.

Process (B_t, C_t^B)**.** We now derive exact results for the size distributions of the processes B_t and C_t^B. So far we have assumed a constant population of healthy cells and studied the asymptotic behaviour of the number of biomarker molecules they shed. Here we first present additional details about this case, providing the exact generating function of C_t^B at a given time t. Later, we show how a derivation similar to that employed for the process (A_t, C_t^A) can be used when B_t is modelled by a critical branching birth-death process.

Constant growth. In the derivation of asymptotic results for our model we already mentioned that if B_t is constant then C_t^B is a branching pure death process with immigration and its exact probability generating function is given by equation (4.3). In our setup we also assume that C_t^B is originally at equilibrium, i.e. that it starts at the expected value of its large time asymptotic distribution. Since we already saw that in this limit C_t^B converges to a Poisson random variable with mean $\frac{\nu_B}{\varepsilon} B_0$, we set $C_0^B = \frac{\nu_B}{\varepsilon} B_0$. With such an initial condition equation (4.3) becomes

$$\mathcal{C}^B(y,t) = \left\{ \left[1 + (y-1)e^{-\varepsilon t}\right] e^{(y-1)(1-e^{-\varepsilon t})} \right\}^{\frac{\nu_B}{\varepsilon} B_0} .$$

Numerical inversion of this function then provides the exact distribution of the process C_t^B at any given time t.

Critical growth. In our model B_t denotes the number of healthy or pre-malignant cells extant at time t and with the ability to shed biomarker molecules in the bloodstream. So far we have assumed that such a number stays constant over time, but in principle we could expect it to fluctuate around a constant average value. For this reason, B_t can be modelled as a critical branching birth-death process, i.e. as a process defined like A_t but with the same birth and death rates $\alpha_B = \beta_B$. Under this assumption, and assuming that the shedding probabilities at cell apoptosis and proliferation are the same for cancerous and healthy cells,

the set of transitions characterizing the two-type process (B_t, C_t^B) are

$$\begin{aligned}
&B \longrightarrow BBC && \text{rate } \alpha_B\,\mu_\alpha\,, && \text{rate } \nu_2'\,,\\
&B \longrightarrow BB && \text{rate } \alpha_B\,(1-\mu_\alpha)\,, && \text{rate } \alpha_B - \nu_2'\,,\\
&B \longrightarrow BC && \text{rate } \nu_1'\,, && \text{rate } \nu_1'\,,\\
&B \longrightarrow C && \text{rate } \alpha_B\,\mu_\beta\,, && \text{rate } \nu_0'\,,\\
&B \longrightarrow \emptyset && \text{rate } \alpha_B\,(1-\mu_\beta)\,, && \text{rate } \alpha_B - \nu_0'\,,\\
&C \longrightarrow \emptyset && \text{rate } \varepsilon\,, && \text{rate } \varepsilon\,.
\end{aligned}$$

The exact probability generating functions for this process can be computed through a derivation similar to that shown for (A_t, C_t^A). Briefly, let us just redefine $\mathcal{P}^*$ as

$$\mathcal{P}^*(x,y,t) = \sum_{m,n\geq 0} x^m y^n \mathbb{P}\left(\left(B_t, C_t^B\right) = (m,n) \mid \left(B_0, C_0^B\right) = *\right)$$

and set

$$\mathcal{B}(x,y,t) = \mathcal{P}^{(1,0)}(x,y,t), \qquad \mathcal{C}'(x,y,t) = \mathcal{P}^{(0,1)}(x,y,t)\,.$$

As biomarker molecules are eliminated by the bloodstream at the same rate regardless of the cell type that shed them, we have $\mathcal{C}'(x,y,t) = \mathcal{C}(x,y,t)$, where $\mathcal{C}(x,y,t)$ is given by equation (4.14). However, the function $\mathcal{B}(x,y,t)$ does not follow straightforwardly from $\mathcal{A}(x,y,t)$ by simply substituting the rates for B_t and taking the limit as $\beta_B \to \alpha_B$. The reason is that in this limit the two linearly independent solutions $h_1(z)$ and $h_2(z)$ of the reduced equation become the same and so one needs to write a general solution to the equation for $f(z)$ in terms of different functions. Adjusting this detail though, the subsequent steps can be repeated and allow to derive an explicit expression for $\mathcal{B}(x,y,t)$. Then, the exact probability generating function for the process (B_t, C_t^B) is equal to

$$\mathcal{P}^{(B_0,C_0^B)}(x,y,t) = \mathcal{B}(x,y,t)^{B_0}\mathcal{C}(x,y,t)^{C_0^B}\,.$$

The marginal generating functions for the two single processes follow from the $y \to 1$ and $x \to 1$ limits, respectively. In particular, for any combination of non-zero shedding rates the former coincides with the probability generating function of a critical branching birth-death process with rate β_B and B_0 initial individuals, as given by equation (1.4). The marginal generating function of C_t^B can instead be inverted numerically to derive the exact probability distribution of the process. However, while the computations just described are feasible, modelling B_t as a

critical branching birth-death process lead to tedious complications and does not practically change the dynamics of the model. These complications are related to the fact that a critical branching process eventually gets extinct with probability one (see Section 1.2.1). Hence, while we are interested in the large time limit distribution of B_t, under these modelling assumptions such a limit converges to a point mass at zero and it is not possible to condition on B_t eventual survival either. On the other hand, since B_t starts with a very large number of cells the time required by B_t to become extinct would be unrealistically long. Indeed, the expected time to extinction for a critical branching process is infinite, and using equation (1.7) we see that if B_t starts with 3×10^8 cells (see Section 4.2.1) the probability it gets extinct in a million years is still around 10^{-4}. For the same principle, the probability that B_t exhibits significant deviations from B_0 within human lifetime is small. To quantify the probability of such fluctuations we can exploit the Chebyshev inequality; using again $\beta_B = 0.1$ and $B_0 = 3 \times 10^8$ this yields $\mathbb{P}(|B_{100y} - 3 \times 10^8| \geq 1.5 \times 10^7) < 0.01$, which says that even waiting 100 years the probability of observing a deviation of at least 2% from the original population size would still be lower than 1%.

Process C_t. Given the independence of biomarker shedding from cancerous and healthy cells, the probability generating function of the process $C_t = C_t^A + C_t^B$ is equal to

$$\mathcal{C}(y,t) = \mathcal{C}^A(y,t)\mathcal{C}^B(y,t)\,,$$

where the two functions on the right hand side are given by equation (4.30) (see also equations (C.3) and (C.6) for special cases) and equation (4.3) (assuming B_t is constant), respectively. Once again, $\mathcal{C}(y,t)$ can be numerically inverted to find the exact distribution of the process C_t.

The probability distributions derived in this section for the processes A_t, B_t and C_t can also be conditioned on A_t eventual survival, or on the size of one of these processes. The form of these distributions, together with expressions for the expected value of the processes and additional information on the implementation of the sampling scheme, are presented in Appendix C.2.

4.5 Discussion

Liquid biopsies combined with deep sequencing of circulating tumor DNA offer new opportunities for cancer early detection and treatment monitoring. How-

ever, whether or not such an approach can also detect small and yet asymptomatic tumors with sufficiently high sensitivity and specificity has not yet been demonstrated. To explore the fundamental biological and mechanistic limitations of ctDNA-based detection tests, we developed a fully stochastic model of tumor growth and ctDNA shedding. In this model, ctDNA fragments are released stochastically by both healthy and cancerous cells and evolve as branching pure death processes. We further assumed that the population of healthy cells remain constant over time, and described the primary tumor growth as a supercritical branching birth-death process. In this framework, we showed that the number of ctDNA fragments present in the bloodstream - or in a blood sample - when the primary tumor is made of a large number of cells follows a Poisson distribution.

By applying this model to an early stage lung cancer cohort, we inferred a mean shedding rate of 1.2×10^{-5} ctDNA whole genome equivalents per cancer cell per day. For tumors with 1 billion cells (≈ 1 cm^3), we calculate that a 15 ml liquid biopsy contains at least two whole genome equivalents of ctDNA with a probability of $\approx 45\%$. Assuming a tumor exhibits three somatic driver point mutations, we expect to find at least two mutant ctDNA fragments for 84% of the patients. In contrast, less than 17% of tumors of size 0.25 cm^3 produce at least two mutant ctDNA fragments per liquid biopsy. An equally sized benign lesion with two driver mutations would be detected with a probability of only 2.2% due to a smaller fraction of cells undergoing apoptosis per unit of time. The same argument explains why fast growing tumors can produce lower levels of ctDNA than slower growing tumors.

While further studies are needed to assess the viability of liquid biopsies and circulating tumor DNA for early cancer detection, our model provides a quantitative framework to explore their potential and limitations. This framework can be applied to different cancer types or extended to more complicated scenarios. For example, it can be used to correlate the tumor burden with the results of repeated biopsies, or with the analysis of blood tests based on a panel of multiple biomarkers shed independently of each other.

Chapter 5

Conclusions

In the introduction of this thesis we highlighted the key role of many mathematical models in the study of cancer evolutionary dynamics. These models are based on different mathematical tools and have collectively been employed to investigate a wide range of biological processes related to the development of a tumor. In particular, the theoretical framework designed by Luria and Delbrück, and its generalization by Lea and Coulson, provide a fitting mathematical description for scenarios where a growing tumor seeds a second population. Inspired by this approach, we employed a similar setting to explore biological aspects underlying metastasis formation and the shedding of cancer biomarkers.

Our models make extensive use of the theory of branching and birth-death processes. In the first two chapters of this thesis we thus presented the main notions in these fields, that we subsequently exploit for the analysis of our models. After discussing preliminary mathematical results, in Chapter 2 we focussed on the probability distributions of first passage times in branching birth-death processes. These random variables can describe how long a cellular population takes to reach a given size, and are thus extremely useful for the study of cancer growth. We presented expressions for the distributions and first moments of hitting times to a finite size, for both noncritical and critical processes. We then investigated the asymptotic limit of these hitting times to large sizes. Such results are particularly relevant for our models, as tumors typically start to show symptoms and become detectable only when made of a large number of cells. The asymptotic distributions of hitting times in noncritical branching birth-death processes are of an extreme value (Gumbel) type. The reason why these times asymptotically behave like the maximum of i.i.d. random variables is not obvious, and we thus presented a mathematical argument to interpret this result.

In Chapter 3 we then discussed our model of cancer recurrence. In this setting we describe the growth of a primary tumor as a deterministic function and assume it seeds metastases at a rate proportional to its size. These secondary lesions are thus initiated as a non-homogeneous Poisson process. Once they are generated, each of them evolves as a supercritical branching birth-death process conditioned on non-extinction. The asymptotic results discussed in the previous chapter allow us to characterize the time taken by each of these metastases to reach a large detectable size. The smallest of these times is then defined as the time to cancer recurrence. In the special case where the primary tumor growth is modelled by an exponential function, we showed that this time follows a Gumbel distribution for the minimum of i.i.d. random variables.

Next, we extended the previous framework to include a scenario where the primary tumor is surgically removed at a given time, modelled by a cut-off time for the primary tumor growth function. By embedding this event, we categorize all metastases into those that become detectable before (synchronous) or after (metachronous) the primary tumor resection. Under these modelling assumptions we derived the probability of several clinically relevant events, in particular that of established but all undetectable metastases, and discussed how the distribution of the relapse time changes when conditioned on these events.

After this qualitative analysis, we tested our model with parameter estimates collected for five different cancer types. The predicted time to cancer recurrence and fraction of patients with synchronous metastases fall in the ranges reported in clinical literature, suggesting that our simple setup is able to capture the main dynamics governing cancer relapse and to provide meaningful information about the time to recurrence.

Finally, in Chapter 4 we introduced a model for cancer biomarker shedding, focussing on circulating tumor DNA. This model can be applied to study liquid biopsies - recently developed screening techniques based on blood tests - to determine their potential to detect cancer early. In this case, we describe the growth of a primary tumor as a supercritical branching birth-death process conditioned on non-extinction. When a cancerous cell dies, it has a given probability of releasing a ctDNA fragment into the circulation. We therefore assumed that the tumor sheds ctDNA at a rate proportional to its size, and that this shedding only happens at cell apoptosis. Under these hypotheses, ctDNA fragments are thus shed

as a Cox process. Since they cannot reproduce and are eliminated by the bloodstream at a fixed rate, we model their evolution as a branching pure death process.

By applying to this setting parameter estimates inferred for lung cancer, we were able to quantify the correlation between tumor burden and ctDNA levels in a plasma sample. We then extended our mathematical framework to allow for biomarker shedding not only at cell apoptosis but also at necrosis or proliferation. Furthermore, while our main results for ctDNA shedding were expressed in their asymptotic forms for large primary tumor size, we derived the probability distribution of biomarker levels at a given time. This model can thus be applied to a wide range of contexts and represents a general mathematical tool to better correlate features of physiological or pathological conditions with the levels of related biomarkers in the blood.

As a whole, the work presented in this thesis stems from non-trivial theoretical results and provides insight into both quantitative and qualitative features of cancer evolution, hence fitting into the context of mathematical contributions to the study of biological processes.

Appendix A

Appendix to Chapter 2

In this appendix we show supplementary material for our discussion of hitting times distributions in branching birth-death processes.

A.1 Hitting times with arbitrary initial condition

Proposition 2.3.2 exploits the result in Theorem 2.3.1 to derive the probability distribution of the hitting time to M in a birth-death process with reflecting boundary at 0 and starting with $1 < n_0 < M$ individuals. Here, we provide a proof based on the Laplace transforms of $\mathbb{P}_0(T_M \leq t)$ and $\mathbb{P}_0(T_{n_0} \leq t)$.

Proof of Proposition 2.3.2. Given the definitions of λ_i and μ_j, Theorem 2.3.1 tells us that

$$\begin{aligned} \mathbb{P}(T_{n_0} \leq t \mid Z_0 = 0) &= \mathbb{P}\left(\sum_{j=1}^{n_0} \mathrm{Exp}(\mu_j) \leq t\right), \\ \mathbb{P}(T_M \leq t \mid Z_0 = 0) &= \mathbb{P}\left(\sum_{i=1}^{M} \mathrm{Exp}(\lambda_i) \leq t\right). \end{aligned} \tag{A.1}$$

Now, let us denote

$$f_{i,j}(t) = \frac{d}{dt}\,\mathbb{P}(T_j \leq t \mid Z_0 = i)$$

the probability density of the first passage time to j starting with i individuals. Then, from equation (A.1) we find

$$\begin{aligned} f_{0,n_0}(t) &= \mu_1 e^{-\mu_1 t} * \cdots * \mu_{n_0} e^{-\mu_{n_0} t}, \\ f_{0,M}(t) &= \lambda_1 e^{-\lambda_1 t} * \cdots * \lambda_M e^{-\lambda_M t}, \end{aligned} \tag{A.2}$$

where the asterisk sign denotes the convolution operation. Additionally, we define

$$\sigma_{i,j}(s) = \mathcal{L}[f_{i,j}](s) = \int_0^\infty f_{i,j}(t)e^{-st}dt$$

as the Laplace transform of the density $f_{i,j}(t)$. By recalling the following properties of Laplace transforms,

$$\mathcal{L}[f * g](s) = \mathcal{L}[f](s) \cdot \mathcal{L}[g](s), \quad \mathcal{L}[ae^{-bt}] = \frac{a}{s+b},$$

equation (A.2) yields

$$\sigma_{0,n_0}(s) = \frac{\mu_1}{s+\mu_1} \cdots \frac{\mu_{n_0}}{s+\mu_{n_0}},$$
$$\sigma_{0,M}(s) = \frac{\lambda_1}{s+\lambda_1} \cdots \frac{\lambda_M}{s+\lambda_M}.$$

On the other hand we also see that $f_{0,M}(t) = f_{0,n_0}(t) * f_{n_0,M}(t)$, which in turn implies $\sigma_{0,n_0}(s) \cdot \sigma_{n_0,M}(s) = \sigma_{0,N}(s)$. The latter identity then leads to

$$\sigma_{n_0,M}(s) = \frac{\sigma_{0,N}(s)}{\sigma_{0,n_0}(s)} = \frac{\lambda_1 \cdots \lambda_M}{\mu_1 \cdots \mu_{n_0}} \frac{\prod_{j=1}^{n_0}(s+\mu_j)}{\prod_{i=1}^{N}(s+\lambda_i)} = C\,\frac{P(s)}{R(s)}.$$

Consider now the rational function $P(s)/R(s)$. Since all the eigenvalues λ_is are distinct (see e.g. [105]) and we are assuming that $n_0 < M$, there exists a unique decomposition of the form

$$\frac{P(s)}{R(s)} = \frac{\prod_{j=1}^{n_0}(s+\mu_j)}{\prod_{i=1}^{M}(s+\lambda_i)} = \sum_{i=1}^{M} \frac{\xi_i}{s+\lambda_i}.$$

Thus, to conclude the proof, we only need to show that the coefficients ξ_i have exactly the form given in equation (2.5). To see this, for any fixed k, $1 \le k \le M$, it is sufficient to multiply both sides of the equation above by $(s+\lambda_k)$,

$$\frac{\prod_{j=1}^{n_0}(s+\mu_j)}{\prod_{\substack{i=1 \\ i\neq k}}^{M}(s+\lambda_i)} = \xi_k + (s+\lambda_k)\sum_{\substack{i=1 \\ i\neq k}}^{M} \frac{\xi_i}{s+\lambda_i}.$$

Then, by substituting $s = -\lambda_k$ we immediately recover for every k the expression given in equation (2.5). □

A.2 First moments

In this section we discuss how to derive the mean and variance of T_M conditional on Ω^M. We consider separately the cases of noncritical and critical branching birth-death processes, and results for finite M and in the asymptotic limit for M large.

A.2.1 Noncritical case

Let us formally define for every $i \geq 1$ the i-th first upward passage time and the i-th interevent time, respectively, as

$$\tau_i = T_{i+1} - T_i, \qquad Y_i = \min\{T_{i+1} - T_i, T_{i-1} - T_i\}\,.$$

The interevent times Y_i of a branching birth-death process are exponentially distributed with parameters $(\alpha + \beta)i$. By denoting as usual $\hat{Z}_n$ the embedded chain, we can also define $\mathbb{1}_{i,j}$ as the indicator function for the event $\{\hat{Z}_1 = j \mid \hat{Z}_0 = i\}$. The first upward passage times τ_i satisfy the following well known recursive relation (see e.g. [105])

$$\tau_i = Y_i + \mathbb{1}_{i,i-1} \cdot (\tau_{i-1} + \tau_i)\,. \tag{A.3}$$

Such a relation will be fundamental for the next proofs.

Finite results

Proof of Proposition 2.3.3. Recall that $\tilde{p}_{i,j}$ denote the one-step transition probabilites of Z_t conditioned on Ω^j. Taking conditional expectations on both sides of equation (A.3) we first find

$$\begin{aligned}\mathbb{E}[\tau_i \mid \Omega_{i+1}] &= \mathbb{E}[Y_i] + \tilde{p}_{i,i-1}(\mathbb{E}[\tau_{i-1} \mid \Omega_{i+1}] + \mathbb{E}[\tau_i \mid \Omega_{i+1}]) \\ &= \frac{1}{\tilde{p}_{i,i+1}}\mathbb{E}[Y_i] + \frac{\tilde{p}_{i,i-1}}{\tilde{p}_{i,i+1}}\mathbb{E}[\tau_{i-1} \mid \Omega_{i+1}] =: a_i\mathbb{E}[Y_i] + b_i\mathbb{E}[\tau_{i-1} \mid \Omega_{i+1}]\,,\end{aligned}$$

by defining $a_i = \frac{1}{\tilde{p}_{i,i+1}}$ and $b_i = \frac{\tilde{p}_{i,i-1}}{\tilde{p}_{i,i+1}}$. Notice that, for every $i \geq 1$, the random variables Y_i does not depend on the event Ω_{i+1} and $\mathbb{E}[\tau_2 \mid \Omega_{i+1}] = \mathbb{E}[Y_1]$.

Therefore, we can iterate the relation above to get

$$\mathbb{E}[\tau_i \mid \Omega_{i+1}] = a_i \mathbb{E}[Y_i] + \sum_{j=1}^{i-1} a_j \left(\prod_{k=j+1}^{i} b_k \right) \mathbb{E}[Y_j] = \sum_{j=1}^{i} \xi_{ij} \mathbb{E}[Y_j] \,, \tag{A.4}$$

where

$$\xi_{ij} = \begin{cases} a_j & \text{if } j = i \,, \\ a_j \left(\prod_{k=j+1}^{i} b_k \right) & \text{if } j < i \,. \end{cases}$$

Now, by plugging in this expression the definitions of a_i and b_i and the formulas for the conditional probabilities given by equation (2.9), we find that

$$\xi_{ij} = \frac{q^i(1+q)(1-q^j)^2}{q^j(1-q^i)(1-q^{i+1})}, \quad \text{for every } j \leq i \,. \tag{A.5}$$

Furthermore, we observe that

$$\mathbb{E}[T_M \mid Z_0 = n_0, \Omega_M] = \sum_{i=n_0}^{M-1} \mathbb{E}[\tau_i \mid \Omega_{i+1}] \,. \tag{A.6}$$

Hence, putting together equations (A.4) to (A.6) and recalling that the interevent times satisfy $Y_i \sim \mathrm{Exp}(i\alpha(1+q))$ for every $i \geq 1$, we find

$$\begin{aligned} \mathbb{E}[T_M \mid Z_0 = n_0, \Omega_M] &= \sum_{i=n_0}^{M-1} \sum_{j=1}^{i} \frac{q^i(1+q)(1-q^j)^2}{q^j(1-q^i)(1-q^{i+1})} \cdot \frac{1}{j\alpha(1+q)} \\ &= \frac{1}{\alpha} \sum_{i=n_0}^{M-1} \frac{q^i}{(1-q^i)(1-q^{i+1})} \sum_{j=1}^{i} \frac{(1-q^j)^2}{jq^j} \,. \end{aligned} \qquad \square$$

Once these expectations are known, they can in turn be used to obtain the conditional variance of T_M. The corresponding proof follows closely the steps made to derive the expectation formula, but require heavier computations, and we will thus only provide a sketch of the demonstration.

Proposition A.2.1. *Let Z_t be a supercritical branching birth-death process with birth rate α and death rate β. Then*

$$\mathrm{Var}(T_M \mid Z_0 = n_0, \Omega_M) = \sum_{i=n_0}^{M-1} \sum_{j=1}^{i} \left[\xi_{ij} \cdot \frac{1}{j^2\alpha^2(1+q)^2} + \eta_{ij} c_j \right] \,, \tag{A.7}$$

where

$$\xi_{ij} = \frac{q^i(1+q)(1-q^j)^2}{q^j(1-q^i)(1-q^{i+1})}, \quad \eta_{ij} = \frac{q^{i+1}(1-q^{i-1})(1-q^j)(1-q^{j+1})}{q^j(1+q)(1-q^i)^2(1-q^{i+1})},$$

and

$$c_j = \left(\mathbb{E}[\tau_{j-1} \mid \Omega_j] + \mathbb{E}[\tau_j \mid \Omega_{j+1}]\right)^2 .$$

Sketch of the proof. Let us define $v_i = \mathrm{Var}(\tau_i \mid \Omega_{i+1})$. Then, by taking variance on both sides of equation (A.3), we find

$$v_i = a_i v_{i-1} + b_i \mathrm{Var}(Y_i) + \tilde{p}_{i,i-1} c_i \,,$$

where a_i and b_i are defined exactly as in the proof of Proposition 2.3.3. Hence, by iterating the relation above and noticing that $v_1 = \mathrm{Var}(Y_1)$, we get

$$v_i = \sum_{j=1}^{i} \xi_{ij} \mathrm{Var}(Y_j) + \sum_{j=1}^{i} \eta_{ij} c_j \,.$$

Here, the first sum follows again in the very same way as for equation (A.3), while the second one comes from the additional term $\tilde{p}_{i,i-1} c_i$ in the recursive relation for v_i. It is then easy to see that

$$\eta_{ij} = \begin{cases} \tilde{p}_{j,j-1} & \text{if } j = i \,, \\ \tilde{p}_{j,j-1} \prod_{k=j+1}^{i} b_k & \text{if } j < i \,, \end{cases}$$

which gives the desired expressions for η_{ij} by just plugging in the definitions of $\tilde{p}_{j,j-1}$, b_k and simplifying the product. The final result follows by summing over i in the range $[n_0, N-1]$. □

Asymptotic results

The asymptotic limits of the quantities we just studied can instead be derived moving from Proposition 2.4.2. Also in this case, we show how to compute the expression for the mean, and then discuss how a similar derivation can be used to obtain the variance of T_M.

Proof of Proposition 2.4.3. Equation (2.16) tells us that the hitting time T_M, conditioned on non-extinction and n_0 initial individuals, is distributed like a mixture of n_0 generalized Gumbel random variables with weights c_j. If we denote f_j the probability densities of these random variables, the density of T_M can be

written as

$$\frac{d}{dt}\,\mathbb{P}(T_M \le t \mid Z_0 = n_0, \Omega_M) = \sum_{j=0}^{n_0-1} c_j f_j(t)\,.$$

Furthermore, by denoting $\mu_j^{(k)} = \int_{-\infty}^{\infty} t^k f_j(t)dt$ the k-th moment of each generalized Gumbel random variable, we see that

$$\mathbb{E}[T_M^k \le t \mid Z_0 = n_0, \Omega_M] = \sum_{j=0}^{n_0-1} c_j \mu_j^{(k)}\,.$$

When $k = 1$, the relation above provides immediately the expression in the statement for $\mathbb{E}[T_M \le t \mid Z_0 = n_0, \Omega_M]$, by simply substituting the parameters of the generalized Gumbel distributions in equation (2.16) and recalling the formulas for their means, given by equation (1.18). □

Using the notation introduced in the previous proof, the variance of T_M can then be written as

$$\mathrm{Var}(T_M \le t \mid Z_0 = n_0, \Omega_M) = \sum_{j=0}^{n_0-1} c_j \mu_j^{(2)} - \left(\mathbb{E}[T_M \le t \mid Z_0 = n_0, \Omega_M]\right)^2$$

and similarly follows by the two results referenced above.

A.2.2 Critical case

We now derive the expression for the mean of T_M conditional on Ω^M in the critical case reported in Proposition 2.3.5.

Proof of Proposition 2.3.5. Let us first define $S_i(x) := \sum_{k=0}^{i-1} x^k$. Then, we have that for every $i \ge 1$, $1-q^i = (1-q)S_i(q)$ and hence we can rewrite equation (2.12) as

$$\mathbb{E}[T_M \mid Z_0 = n_0, \Omega_M] = \frac{1}{\alpha} \sum_{i=n_0}^{M-1} \frac{q^i}{S_i(q)S_{i+1}(q)} \sum_{j=1}^{i} \frac{S_j^2(q)}{jq^j}.$$

Then, since $S_i(1) = i$ for every i, the statement follows by simply substituting $q = 1$ in the expression above. □

The corresponding variance of T_M can be similarly obtained from equation (A.7). However, such a derivation involves long and tedious computations that we do not present here. For completeness, we report its expression, which reads

$$\frac{1}{2\alpha^2}\left\{\frac{1}{M} - \frac{1}{n_0} + \frac{1}{3}(H_{M-1} - H_{n_0-1}) + \frac{1}{6}[M(M+1) - n_0(n_0+1)]\right\}, \qquad \text{(A.8)}$$

where $H_k = \sum_{i=1}^{k} i^{-1}$ denotes the k-th harmonic number.

The asymptotic limits of this mean and variance as M gets large follow straightforwardly from their finite expressions, and are shown in equation (2.17).

A.3 Asymptotic distributions

In this section we present the steps necessary to derive the asymptotic distribution of the first passage time to M, as M gets large, conditioned on Ω^M and $n_0 \geq 1$ initial individuals.

A.3.1 Noncritical case

Theorem 2.4.1 provides the asymptotic distribution of the first passage time to a large size M for a supercritical branching birth-death process conditioned on Ω^M and one initial individual. The proof we presented for such theorem relies on a martingale argument, that can be extended to derive a similar result for processes starting with $n_0 > 1$ individuals. To this purpose, we first show a few auxiliary results.

Let Z_t be a supercritical branching birth-death process with birth rate α, death rate β and starting with $n_0 \geq 1$ individuals. Moreover, denote $\lambda = \alpha - \beta > 0$ its net growth rate and $\Omega_\infty = \{Z_t > 0 \text{ for all } t\}$ its eventual survival event, for which we recall that $\mathbb{P}(\Omega_\infty \mid Z_0 = 1) = 1 - \beta/\alpha = 1 - q < 1$ (see equation (1.8)). Following the steps shown in Section 1.2.1, we see that Z_t divided by its mean, $\mathbb{E}[Z_t] = n_0 e^{\lambda t}$, is a non-negative martingale. Such a process thus converges to a limit non-negative random variable $W_{(n_0)}$. Repeating now the argument used in the proof of Theorem 2.4.1 we can write

$$\lim_{t\to\infty} \frac{Z_t}{n_0 e^{\lambda t}} \overset{a.s.}{=} W_{(n_0)} \iff \lim_{M\to\infty} \frac{M}{n_0 e^{\lambda T_M}} \overset{a.s.}{=} W_{(n_0)} \iff \lim_{M\to\infty} \frac{M}{e^{\lambda T_M}} \overset{a.s.}{=} Y_{(n_0)}\,,$$

where $Y_{(n_0)} := n_0 W_{(n_0)}$. The probability distribution of the limit random variable $Y_{(n_0)}$ conditioned on non-extinction is provided by the following.

Proposition A.3.1. *Let $Y_{(n_0)}$ be defined as above. Then its cumulative distribution function conditioned on Ω_∞ is equal to*

$$\mathbb{P}(Y_{(n_0)} \leq t \mid \Omega_\infty) = 1 - \sum_{j=0}^{n_0-1} c_j \left(e^{-(1-q)t} \sum_{i=0}^{n_0-j-1} \frac{[(1-q)t]^i}{i!} \right), \tag{A.9}$$

where $c_j := \dfrac{\binom{n_0}{j} q^j (1-q)^{n_0-j}}{1-q^{n_0}}$ *for every* $j = 0, \dots, n_0 - 1$.

Proof. When $Z_0 = n_0$, the branching property (see Section 1.2.1) yields

$$Z_t = \sum_{i=1}^{n_0} Z_t^{(i)} ,$$

where $Z_t^{(i)}$ are the independent processes generated by the initial n_0 individuals. In particular, each $Z_t^{(i)}$ is an i.i.d. copy of a supercritical branching birth-death process with birth rate α, death rate β and starting with one individual. Hence, for every $i = 1, \dots, n_0$ we have

$$\lim_{t\to\infty} \frac{Z_t^{(i)}}{e^{\lambda t}} \overset{a.s.}{=} Y_{(1)}^{(i)} ,$$

where $Y_{(1)}^{(i)}$ are n_0 i.i.d. copies of the limit random variables W defined by equation (1.11). Then, we can write

$$Y_{(n_0)} \overset{a.s.}{=} \lim_{t\to\infty} \frac{Z_t}{e^{\lambda t}} = \sum_{i=1}^{n_0} \left[\lim_{t\to\infty} \frac{Z_t^{(i)}}{e^{\lambda t}} \right] \overset{a.s.}{=} \sum_{i=1}^{n_0} Y_{(1)}^{(i)} . \tag{A.10}$$

If we know how many of the initial processes survive, we can then use the explicit distribution of $Y_{(1)}^{(i)}$ - given by equation (1.12) - to derive that of $Y_{(n_0)}$. Hence, let us define $\Omega_0^{(j)} = \{Z_t^{(j)} = 0 \text{ for some } t\}$ and the random variable B as

$$B = \sum_{j=0}^{n_0} \mathbb{1}_{\Omega_0^{(j)}} ,$$

where $\mathbb{1}$ denotes the indicator function of an event. Intuitively, B represents the number of the initial processes $Z_t^{(i)}$ which will eventually die out. As we know that these processes independently get extinct with the same probability q, we have that $B \sim \mathrm{Bin}(n_0, q)$. By averaging over B, we then find

$$\mathbb{P}(Y_{(n_0)} \le t) = \sum_{j=0}^{n_0} \mathbb{P}(B = j)\, \mathbb{P}(Y_{(n_0)} \le t \mid B = j) . \tag{A.11}$$

Since the process Z_t gets extinct if and only if all of the initial processes $Z_t^{(i)}$ die

out, equation (A.11) can be further expanded as

$$P(Y_{(n_0)} \leq t) = q^{n_0} + \sum_{j=0}^{n_0-1} \mathbb{P}(B = j)\mathbb{P}(Y_{(n_0)} \leq t|B = j)\,. \tag{A.12}$$

At the same time, since Ω_0 and Ω_∞ constitute a partition of the sample space we also have

$$\begin{aligned}\mathbb{P}(Y_{(n_0)} \leq t) &= \mathbb{P}(Y_{(n_0)} \leq t \mid \Omega_0) \cdot \mathbb{P}(\Omega_0) + \mathbb{P}(Y_{(n_0)} \leq t \mid \Omega_\infty) \cdot \mathbb{P}(\Omega_\infty) \\ &= q^{n_0} + (1 - q^{n_0})\mathbb{P}(Y_{(n_0)} \leq t \mid \Omega_\infty)\,.\end{aligned} \tag{A.13}$$

Therefore, by combining equations (A.12) and (A.13) we immediately find

$$\mathbb{P}(Y_{(n_0)} \leq t \mid \Omega_\infty) = \frac{1}{1 - q^{n_0}} \sum_{j=0}^{n_0-1} \mathbb{P}(B = j)\,\mathbb{P}(Y_{(n_0)} \leq t \mid B = j)\,. \tag{A.14}$$

Now, the term $\mathbb{P}(Y_{(n_0)} \leq t|B = j)$ is the distribution of the limit random variable $Y_{(n_0)}$ when $n_0 - j$ of the initial processes survive. By combining equations (1.12) and (A.10), we thus find that this distribution coincides with that of a sum of $n_0 - j$ i.i.d. exponential random variables with parameter $1 - q$. Equivalently, it is an Erlang distribution with parameters $n_0 - j$ and $1 - q$, and is thus given by (see e.g. [59])

$$\mathbb{P}(Y_{(n_0)} \leq t|B = j) = \left(1 - e^{-(1-q)t} \sum_{i=0}^{n_0-j-1} \frac{[(1 - q)t]^i}{i!}\right)\,.$$

The statement of the proposition then follows by plugging back into the expression above and that for the binomial mass function of B. □

Equation (A.9) can now be used to prove Proposition 2.4.2 and derive a formula for the asymptotic limit of T_M distribution conditioned on $n_0 \geq 1$ initial individuals.

Proof of Proposition 2.4.2. The cumulative distribution function of the k-th generalized Gumbel distribution is given by equation (1.17). By substituting such expression in equation (2.16), we see that the statement of the proposition is equivalent to

$$\mathbb{P}_{n_0}(T_M \leq t \mid \Omega_\infty) \sim \sum_{j=0}^{n_0-1} c_j \left(e^{-(1-q)Me^{-\lambda t}} \sum_{i=0}^{n_0-j-1} \frac{\left[(1 - q)Me^{-\lambda t}\right]^i}{i!}\right)\,, \tag{A.15}$$

asymptotically for large M. Now, following the same steps we used for the proof of Theorem 2.4.1, for large M we find

$$\begin{aligned}\mathbb{P}(T_M \leq t \mid Z_0 = n_0, \Omega_\infty) &= \mathbb{P}(Me^{-\lambda T_M} \geq Me^{-\lambda t} \mid Z_0 = n_0, \Omega_\infty) \\ &\sim 1 - \mathbb{P}(Y_{(n_0)} \leq Me^{-\lambda t} \mid \Omega_\infty)\,.\end{aligned}$$

The last expression can then be expanded using equation (A.9) and provides exactly equation (A.15). □

A.3.2 Critical case

As M gets large, the asymptotic distribution of T_M for a critical branching birth-death process conditioned on non-extinction is determined by the corresponding limit of the zeros of the associated Laguerre polynomials $L_M^{(1)}(x)$. Here we thus present classical results about the asymptotics of these zeros, and then show how to use them to prove Theorem 2.4.4.

Proposition A.3.2. *Let $\lambda_{M,i}$ be the M zeros of the associated Laguerre polynomial of order 1, $L_M^{(1)}(x)$, in increasing order. Denote j_i the positive zeros of the Bessel function of the first kind, $J_1(x)$, also in increasing order. Then we have*

$$\lim_{M\to\infty} (4M+4)\lambda_{M,i} = j_i^2\,.$$

For a proof of this result, see for instance [73] and the references therein. The asymptotic distribution of T_M can then be obtained as follows.

Proof of Theorem 2.4.4. Proposition A.3.2 implies that as M gets large, the zeros of $L_M^{(1)}(x)$ satisfy $\lambda_{M,i} \sim \dfrac{j_i^2}{4M+4}$. Hence, in the same limit we also have $\alpha\lambda_{M-1,i} \sim \dfrac{\alpha j_i^2}{4M}$. The statement thus follows by combining this asymptotic result with Proposition 2.3.4. □

A.4 Extreme value interpretation

In this section we provide auxiliary proofs to show the symmetry between the asymptotic conditional first passage times distributions of a branching supercritical birth-death process Z_t and the reversed process Z_t^*.

Proof of Lemma 2.4.7. Since the last exit time R_M^* is a non-negative random variable, to show that it converges to zero in distribution it is sufficient to show

that its expectation tends to zero, thanks to Markov's inequality. First define

$$K_M^* = \#\text{ of } Z_t^* \text{ visits to } M\,,$$
$$\Theta_M^* = \inf\{t > 0 : Z_t^* = M, Z_s^* \neq M \text{ for some } 0 < s < t\}\,,$$

and let us denote $\mathbb{E}_M^* = \mathbb{E}[\,\cdot \mid Z_0^* = M]$. Then, we can write

$$\mathbb{E}_M^*[R_M^* \mid \Omega_M^*] = \mathbb{E}_M^* \left[\sum_{j=1}^{K_M^*-1} \Theta_M^* \,\middle|\, \Omega_M^* \right],$$

and from Wald's equality this implies

$$\mathbb{E}_M^*[R_M^* \mid \Omega_M^*] = \mathbb{E}_M^*[K_M^* - 1 \mid \Omega_M^*]\mathbb{E}_M^*[\Theta_M^* \mid \Omega_M^*]\,. \tag{A.16}$$

Now, since Z_t^* is a subcritical birth-death process, it will get extinct with probability one and hence all the states except 0 are transient. It follows in particular that almost surely

$$1 \leq \mathbb{E}_M^*[K_M^* - 1 \mid \Omega_M^*] < \infty\,. \tag{A.17}$$

To prove our result we then need to show that $\mathbb{E}_M^*[\Theta_M^* \mid \Omega_M^*]$ goes to zero. Consider the embedded chain $\hat{Z}_n^*$ and define

$$\sigma_M^* = \inf\{n > 0 : \hat{Z}_n^* = M\}\,.$$

If Z_t^* starts at with M individuals, σ_M^* represents the number of transitions that the process goes through before it gets back to size M for the first time. Furthermore, if the process does go back to M at a certain time before extinction, this time will be distributed as the sum of σ_M^* exponential inter-event times. Now, we also define the following

$$A = \left\{ \omega : \hat{Z}_n^*(\omega) > \frac{M}{2} \;\text{ for all }\; 0 \leq n \leq \sigma_M^*(\omega) \right\}.$$

When Z_t starts at M, the above represents the event that Z_t does not reach sizes smaller than or equal to $M/2$ before it gets back to M for the first time. It is then easy to check that

$$\mathbb{P}_M^*(A \mid \Omega_M^*) \geq \left(\frac{\alpha\beta}{\alpha + \beta} \right)^{M/2} = e^{-cM}\,, \tag{A.18}$$

where $c = \frac{1}{2}\log\left(\frac{\alpha\beta}{\alpha+\beta}\right)$. Let us denote n_i, where $i = 1, \ldots, \sigma_M^*$, all the states that the process Z_t^* goes through in its trajectory back to M. We can then write

$$\mathbb{E}_M^*[\Theta_M^* \mid \Omega_M^*] = \mathbb{E}_M^* \left[\sum_{i=1}^{\sigma_M^*} \mathrm{Exp}(n_i(\alpha+\beta)) \,\middle|\, \Omega_M^* \right] = \mathbb{E}_M^* \left[\sum_{i=1}^{\sigma_M^*} \frac{1}{n_i(\alpha+\beta)} \,\middle|\, \Omega_M^* \right] .$$

By averaging the last expression over A and applying equation (A.18), we find that $\mathbb{E}_M^*[\Theta_M^* \mid \Omega_M^*]$ is bounded above by

$$\mathbb{E}_M^*\left[\sigma_M^* \mid \Omega_M^*\right] \cdot \frac{2}{M(\alpha+\beta)} \cdot (1 - e^{-cM}) + \mathbb{E}_M^*\left[\sigma_M^* \mid \Omega_M^*\right] \cdot \frac{1}{1 \cdot (\alpha+\beta)} \cdot e^{-cM} .$$

Finally, as we observed for the number of visits K_M^*, the fact that M is a transient state implies that $\mathbb{E}_M^*\left[\sigma_M^* \mid \Omega_M^*\right] < \infty$ as well and so the bound above yields

$$\mathbb{E}_M^*[\Theta_M^* \mid \Omega_M^*] \overset{M\to\infty}{\longrightarrow} 0 . \tag{A.19}$$

The statement then follows by joining together equations (A.16), (A.17) and (A.19) and using Markov's inequality. □

Appendix B

Appendix to Chapter 3

In this appendix we provide additional details about the mathematical features of our model for cancer recurrence.

B.1 Scaled relapse time distribution

In Section 3.2 we derived the general expression for the relapse time distribution, which can be written as

$$\mathbb{P}(\tau \leq t) = 1 - e^{-\nu(1-q)\int_0^t n(s)G(t-s)ds} = 1 - e^{-\nu(1-q)\int_0^t n(s)e^{-(1-q)Me^{-\lambda(t-s)}}ds},$$

by combining equations (3.2) to (3.4). Here we show how to scale the detectable size M out of the previous expression, so to split the distribution into a deterministic part and a stochastic term. Let us focus on the integral

$$\int_0^t n(s)e^{-(1-q)Me^{-\lambda(t-s)}}ds$$

and apply two changes of variables: first we use $s \to t - s$ and transform it into

$$\int_0^t n(t-s)e^{-(1-q)Me^{-\lambda s}}ds\,,$$

then we apply $s \to s - \frac{1}{\lambda}\log M$ to find

$$\int_{-\frac{1}{\lambda}\log M}^{t-\frac{1}{\lambda}\log M} n\left(t - s - \frac{1}{\lambda}\log M\right)e^{-(1-q)e^{-\lambda s}}ds\,.$$

By plugging this expression back into equation (3.4) we observe that

$$\mathbb{P}\left(\tau - \frac{1}{\lambda}\log M \leq t\right) = 1 - e^{-\nu(1-q)\int_{-\frac{1}{\lambda}\log M}^{t} n(t-s)e^{-(1-q)e^{-\lambda s}}ds}\,.$$

As M tends to infinity we obtain

$$\tau - \frac{1}{\lambda}\log M \xrightarrow[M\to\infty]{d} \bar{\tau}\,,$$

where

$$\mathbb{P}(\bar{\tau} \le t) = 1 - e^{-\nu(1-q)\int_{-\infty}^{t} n(t-s)e^{-(1-q)e^{-\lambda s}}ds}\,.$$

From the last two equations we also see that asymptotically as $M \to \infty$

$$\mathbb{E}[\tau] \sim \frac{1}{\lambda}\log M + C, \qquad C = \mathbb{E}[\bar{\tau}]\,. \tag{B.1}$$

B.2 Explicit results for exponential primary growth

Two commonly employed growth functions for primary tumors are the exponential $n_e(t) = e^{\delta t}$ and the logistic $n_l(t) = \frac{Ke^{\delta t}}{K+e^{\delta t}-1}$ ones (see e.g. [156]). A logistic growth implies that the primary tumor has a carrying capacity K. During the first stages of its development $n_l(t)$ follows the same exponential trajectory of $n_e(t)$ and then approaches a constant growth as it gets closer to size K. As the carrying capacity is typically large, this slowdown for $n_l(t)$ happens around $\hat{t} = \log(K)/\delta$. The differences between the results provided by these two growths functions thus depend on the probability of metastases being initiated by time $\hat{t}$, i.e. $\mathbb{P}(K_{\hat{t}} \ge 1) \approx 1 - e^{-\frac{\nu(1-q)K}{\delta}}$. Hence, if

$$\frac{\nu(1-q)K}{\delta} \gg 1\,, \tag{B.2}$$

metastases likely establish in the first stages of the primary growth, i.e. when $n_l(t) \approx n_e(t)$. Otherwise, metastases are initiated late in the primary evolution, when the two growth functions are substantially different. This feature is visualized in Figure 3.1, where τ densities for a logistic growth are shown to converge to the exponential ones as ν increases and the other parameters are fixed.

Using the parameter values from Table 3.2, however, we observe that the condition in equation (B.2) is satisfied for all cancer types considered. In other words, our estimates for ν, q, K and δ yield no difference between exponential and logistic growth functions. In light of this, here we study in greater detail the results obtained with $n_e(t)$.

Scaled relapse time. When $n(t) = n_e(t) = e^{\delta t}$, the relapse time distribution has an expression in terms of special functions. To show this, let us consider the

distribution of the scaled relapse time $\overline{\tau}$ as given by equation (3.5) and focus on the integral

$$\int_{-\infty}^{t} n(t-s)e^{-(1-q)e^{-\lambda s}}ds = e^{\delta t}\int_{-\infty}^{t} e^{-(1-q)e^{-\lambda s}-\delta s}ds\,.$$

This can be equivalently written as

$$e^{\delta t}\int_{-\infty}^{t} \frac{1}{(1-q)^{\delta/\lambda}}\left[(1-q)e^{-\lambda s}\right]^{\delta/\lambda} e^{-(1-q)e^{-\lambda s}}ds\,.$$

The last expression then suggests the change of variable $x = (1-q)e^{-\lambda s}$, that in turn leads to

$$\frac{e^{\delta t}}{\lambda(1-q)^{\delta/\lambda}}\int_{(1-q)e^{-\lambda t}}^{\infty} x^{\frac{\delta}{\lambda}-1}e^{-x}dx = \frac{e^{\delta t}}{\lambda(1-q)^{\delta/\lambda}}\Gamma\left(\frac{\delta}{\lambda},(1-q)e^{-\lambda t}\right)\,,$$

where the function Γ above denotes the incomplete upper gamma function

$$\Gamma(a,t) = \int_{t}^{\infty} x^{a-1}e^{-x}dx\,.$$

The scaled relapse time distribution for $n(t) = e^{\delta t}$ is thus given by

$$\mathbb{P}(\bar{\tau} \leq t) = 1 - e^{-\frac{\nu(1-q)^{1-\frac{\delta}{\lambda}}e^{\delta t}}{\lambda}\Gamma\left(\frac{\delta}{\lambda},(1-q)e^{-\lambda t}\right)}\,. \tag{B.3}$$

Since $\Gamma(1,t) = e^{-t}$, for $\lambda = \delta$ this simplifies to

$$\mathbb{P}(\bar{\tau} \leq t) = 1 - e^{-\frac{\nu}{\lambda}e^{-(1-q)e^{-\lambda t}+\lambda t}}\,.$$

Small initiation limit. While the initiation rate can vary significantly across different cancer types, ν is typically orders of magnitude smaller then all other parameters. Hence, we now investigate τ distribution in the $\nu \to 0$ limit. Let us first consider the result given by equation (B.3) for the scaled time to recurrence $\overline{\tau}$ and write it as

$$\begin{aligned}\mathbb{P}(\bar{\tau} \leq t) &= 1 - e^{-\frac{\nu(1-q)^{1-\frac{\delta}{\lambda}}e^{\delta t}}{\lambda}\left(\Gamma\left(\frac{\delta}{\lambda}\right)-\phi(t)\right)} \\ &= 1 - e^{-\frac{\nu(1-q)^{1-\frac{\delta}{\lambda}}\Gamma\left(\frac{\delta}{\lambda}\right)e^{\delta t}}{\lambda}}e^{\frac{\nu(1-q)^{1-\frac{\delta}{\lambda}}\phi(t)e^{\delta t}}{\lambda}}\,,\end{aligned} \tag{B.4}$$

where

$$\phi(t) = \int_0^{(1-q)e^{-\lambda t}} s^{\frac{\delta}{\lambda}-1} e^{-s} ds < \int_0^{(1-q)e^{-\lambda t}} s^{\frac{\delta}{\lambda}-1} ds = \frac{\lambda}{\delta}(1-q)^{\frac{\delta}{\lambda}} e^{-\delta t}\,.$$

Notice that the second exponential factor in equation (B.4) is bounded below by 1 and above by $e^{\frac{\nu(1-q)}{\delta}}$. Therefore, as $\nu \to 0$, the distribution of $\bar{\tau}$ asymptotically converges to

$$\mathbb{P}(\bar{\tau} \leq t) \sim 1 - \exp\left(-\frac{\nu(1-q)^{1-\frac{\delta}{\lambda}}\Gamma\left(\frac{\delta}{\lambda}\right)e^{\delta t}}{\lambda}\right).$$

Equivalently, for small initiation rates the scaled relapse time $\bar{\tau}$ asymptotically follows a Gumbel distribution for the minimum (see Section 1.2.2)

$$\bar{\tau} \sim \mathrm{Gumb}_{\min}\left(-\frac{1}{\delta}\log\frac{\nu(1-q)^{1-\frac{\delta}{\lambda}}\Gamma\left(\frac{\delta}{\lambda}\right)}{\lambda}, -\frac{1}{\delta}\right).$$

Mean relapse time. By combining the last result with equation (1.21) and (B.1) we find that

$$\mathbb{E}[\tau] \approx \frac{1}{\lambda}\log M + \frac{1}{\delta}\log\frac{\delta}{\nu} + C\,,$$

where $C = -\frac{1}{\delta}\left(\log\frac{\delta(1-q)^{1-\frac{\delta}{\lambda}}\Gamma\left(\frac{\delta}{\lambda}\right)}{\lambda} + \gamma_E\right)$. Intuitively, the time to relapse is likely to be determined by one of the first established metastases. Given the simple dependence of $\mathbb{E}[\tau]$ on M and ν, we now compare it with the mean time to detectability of the first metastasis, $\mathbb{E}[\tau_1]$. Let us first recall that $\tau_1 = \sigma_1 + \Theta_1$ is equal to the sum of the first initiation time and the hitting time to M. As $\nu \to 0$, the distribution of the first arrival σ_1, given in general by $1 - e^{-a_t}$, converges to a reverse Gumbel with parameters $\frac{1}{\delta}\log\frac{\delta}{\nu(1-q)}$ and $-\frac{1}{\delta}$. This implies in particular that

$$\mathbb{E}[\sigma_1] = \frac{1}{\delta}\left(\log\frac{\delta}{\nu(1-q)} - \gamma_E\right) = \frac{1}{\delta}\log\frac{\delta}{\nu} + C_1\,,$$

where $C_1 = -\frac{\log(1-q)+\gamma_E}{\delta}$. Moreover, the hitting times Θ_i follow the Gumbel distribution $G(t)$ - see equation (3.2) - and hence

$$\mathbb{E}[\Theta_i] = \frac{1}{\lambda}\log M - C_2\,,$$

for every i, where $C_2 = -\frac{\log(1-q)+\gamma_E}{\lambda}$. Joining the last two results we get

$$\mathbb{E}[\tau_1] = \mathbb{E}[\sigma_1 + \Theta_1] = \frac{1}{\lambda}\log M + \frac{1}{\delta}\log\frac{\delta}{\nu} + \tilde{C}\,, \tag{B.5}$$

where $\tilde{C} = C_1 - C_2$. By comparing equation (B.5) with the expression for $\mathbb{E}[\tau]$, we notice indeed the same M and ν dependence, but the constants C and $\tilde{C}$ have different analytical forms.

Numerical computation. Finally, all the plots and computations reported in this paper have been performed on Matlab R2018b. The lines of code below provide an efficient way (in the example for the exponential case) to calculate the relapse time distribution given by equation (3.4) for a vector of times `tspan`.

```
n = @(t)(exp(delta*t));
G = @(t)(exp(-(1-q)*M*exp(-lambda*t)));
F = @(t)(1-exp(-nu*(1-q)*integral(@(s)(n(s).*G(t-s)),0,t,'ArrayValued',true)));
x = arrayfun(@(t)F(t),tspan);
```

Appendix C

Appendix to Chapter 4

In this appendix we show supplementary information for the model discussed in Chapter 4. Following the structure used there, we first present material related to the asymptotic distributions of the processes considered.

C.1 Asymptotic means and variances

Here we briefly summarize the expected values and variances of the processes A_t and C_t in the asymptotic limits considered. These moments follow straightforwardly from the probability distributions derived in Chapter 4, and can be later compared with their exact form. The process A_t is a supercritical branching birth-death process with net growth rate λ_A, extinction probability q_A and such that $A_0 = 1$. While at time $T_M = \inf\{A_t = M\}$ the process A_t is clearly equal to M, referring back to Section 1.2.1 we see that for large t

$$\mathbb{E}[A_t] = e^{\lambda_A t}\,, \quad \mathrm{Var}(A_t) \sim \frac{1+q_A}{1-q_A}\, e^{2\lambda_A t}\,,$$

and

$$\mathbb{E}\left[A_t \mid \Omega_\infty^A\right] \sim \frac{e^{\lambda_A t}}{1-q_A}\,, \quad \mathrm{Var}(A_t \mid \Omega_\infty^A) \sim \frac{1+2q_A}{(1-q_A)^2}\, e^{2\lambda_A t}\,,$$

where Ω_∞^A, denoting the event of A_t eventual survival, is such that $\mathbb{P}(\Omega_\infty^A) = 1-q_A$.

For the process C_t^A, given that $(A_0, C_0^A) = (1, 0)$, equations (4.2) and (4.8) yield

$$\mathbb{E}\left[C_{T_M}^A \mid \Omega_\infty^A\right] \sim \frac{\nu_A\, M}{\varepsilon + \lambda_A}\,, \quad \mathrm{Var}\left(C_{T_M}^A \mid \Omega_\infty^A\right) \sim \frac{\nu_A\, M}{\varepsilon + \lambda_A}\,,$$

and

$$\mathbb{E}\left[C_t^A \mid \Omega_\infty^A\right] \sim \frac{\alpha_A \nu_A}{\lambda_A(\varepsilon+\lambda_A)}\, e^{\lambda_A t}\,,$$
$$\mathrm{Var}\left(C_t^A \mid \Omega_\infty^A\right) \sim \frac{\alpha_A \nu_A}{\lambda_A(\varepsilon+\lambda_A)}\, e^{\lambda_A t}\left(\frac{\alpha_A \nu_A}{\lambda_A(\varepsilon+\lambda_A)}\, e^{\lambda_A t}+1\right),$$

for a large primary tumor size and for a large time, respectively.

Next, while B_t is assumed constant, the process C_t^B exhibits the same large primary tumor size and large time asymptotic behaviour. For these, conditional on $C_0^B = \frac{\nu_B}{\varepsilon} B_0$, we have

$$\lim_{M\to\infty} \mathbb{E}\left[C_{T_M}^B\right] = \lim_{t\to\infty} \mathbb{E}[C_t^B] = \frac{\nu_B}{\varepsilon} B_0\,,$$
$$\lim_{M\to\infty} \mathrm{Var}\left(C_{T_M}^B\right) = \lim_{t\to\infty} \mathrm{Var}\left(C_t^B\right) = \frac{\nu_B}{\varepsilon} B_0\,.$$

Joining these results together and applying the same initial conditions, for the process C_t we eventually find

$$\mathbb{E}\left[C_{T_M} \mid \Omega_\infty^A\right] \sim \frac{\nu_A\, M}{\varepsilon+\lambda_A} + \frac{\nu_B}{\varepsilon} B_0\,, \quad \mathrm{Var}\left(C_{T_M} \mid \Omega_\infty^A\right) \sim \frac{\nu_A\, M}{\varepsilon+\lambda_A} + \frac{\nu_B}{\varepsilon} B_0\,,$$

for large M, and

$$\mathbb{E}\left[C_t \mid \Omega_A^\infty\right] \sim \frac{\alpha_A \nu_A}{\lambda_A(\varepsilon+\lambda_A)}\, e^{\lambda_A t} + \frac{\nu_B}{\varepsilon} B_0\,,$$
$$\mathrm{Var}\left(C_t \mid \Omega_A^\infty\right) \sim \frac{\alpha_A \nu_A}{\lambda_A(\varepsilon+\lambda_A)}\, e^{\lambda_A t}\left(\frac{\alpha_A \nu_A}{\lambda_A(\varepsilon+\lambda_A)}\, e^{\lambda_A t}+1\right) + \frac{\nu_B}{\varepsilon} B_0\,.$$

for large t.

C.2 Exact results

In this section we provide additional details about the exact distributions of the processes A_t, B_t and C_t.

C.2.1 Special cases of multiple shedding dynamics

Here we present the special forms that the exact distributions derived in Section 4.4.2 when one or more of the parameters $\mu_\beta, \nu_1, \mu_\alpha$ are zero. In general, among the nonzero shedding rates the one with the highest index determines the special functions involved in the generating function $\mathcal{A}(x, y, t)$. As we showed

these are confluent Heun functions when $\nu_2 > 0$, while we will see that they become confluent hypergeometric functions when $\nu_2 = 0$ and $\nu_1 > 0$ and Bessel functions of the first kind when $\nu_2 = 0, \nu_1 = 0$ and $\nu_0 > 0$. The remaining cases follow by straightforward substitution from these three. Hence, we will now provide a sketch of how the previous derivation can be adapted to the cases $\nu_2 = 0, \nu_1 > 0$ and $\nu_2 = 0, \nu_1 = 0, \nu_0 > 0$. The steps up until equation (4.17) are valid for all scenarios, so that will be the starting point of the adapted derivations. Also notice that for all these cases the marginal probability generating function $\mathcal{A}(y,t)$ remains unchanged, as the evolution of the process A_t is not affected by its shedding activity.

Case $\mu_\alpha = 0, \nu_1 > 0$

When $\mu_\alpha = 0$ we have $b_2 = 0$. Hence, equation (4.17) becomes

$$\partial_s^2 Y - \frac{b_1 s + a_1 - 1}{s} \partial_s Y + \frac{a_2(a_0 + b_0 s)}{s^2} Y = 0\,.$$

By seeking directly a solution of the form $Y(s) = s^{-a_2} f(s)$ and then applying the change of variables $z = b_1 s$, the equation above reduces to the standard form of Kummer's equation [3]

$$z f''(z) + (\gamma - z) f'(z) - \omega f(z) = 0\,, \tag{C.1}$$

where

$$\begin{aligned} \gamma &= a_0 - a_2 + 1\,, \\ \omega &= -a_2 \left(\frac{b_0}{b_1} + 1 \right)\,. \end{aligned}$$

Including again only one integrating constant D, its solution can be expressed in terms of the confluent hypergeometric function ${}_1F_1$ as

$$f(z) = {}_1F_1(\omega, \gamma, z) + D z^{1-\gamma} {}_1F_1(\omega - \gamma + 1, 2 - \gamma, z)\,.$$

Substituting back we immediately find an expression for the solution $Y(s)$ in terms of

$$h_1(s) = {}_1F_1(\omega, \gamma, b_1 s), \quad h_2(s) = {}_1F_1(\omega - \gamma + 1, 2 - \gamma, b_1 s)\,.$$

Furthermore, since

$$\frac{\partial}{\partial z}{}_1F_1(a,b,z) = \frac{a}{b}{}_1F_1(a+1,b+1,z)\,,$$

the derivative of $Y(s)$ can be computed in terms of the functions h_1, h_2, g_1 and g_2, where

$$g_1(s) = {}_1F_1(\omega+1,\gamma+1,b_1 s), \quad g_2(s) = {}_1F_1(\omega-\gamma+2,3-\gamma,b_1 s)\,.$$

Joining these results together we find an expression for $X(s)$, in terms of the same functions. This expression still depends on the integrating constant D, which is determined by the initial condition $X(x,y,s=1) = a_2 x$. Then, multiplying $X(x,y,s)$ by $\frac{s}{a_2}$ and substituting $s = e^{-\varepsilon t}$ we eventually find

$$\mathcal{A}(x,y,t) = \frac{e^{\lambda t}K_1(x,y)\phi_2(y,t) - K_2(x,y)\phi_1(y,t)}{e^{\lambda t}K_1(x,y)\psi_1(y,t) - K_2(x,y)\psi_2(y,t)}\,, \tag{C.2}$$

where

$$\begin{aligned}
K_1(x,y) &= (\alpha-\beta+\varepsilon)(1-x)h_1(y,0) - (\nu_1+\nu_0)(y-1)g_1(y,0)\,,\\
K_2(x,y) &= (\alpha-\beta-\varepsilon)(\beta-\alpha x)h_2(y,0) + (\beta\nu_1+\alpha\nu_0)(y-1)g_2(y,0)\,,\\
\phi_1(y,t) &= (\alpha-\beta+\varepsilon)h_1(y,t) - (\nu_1+\nu_0)(y-1)e^{-\varepsilon t}g_1(y,t)\,,\\
\phi_2(y,t) &= \beta(\alpha-\beta-\varepsilon)h_2(y,t) + (\beta\nu_1+\alpha\nu_0)(y-1)e^{-\varepsilon t}g_2(y,t)\,,\\
\psi_1(y,t) &= (\alpha-\beta+\varepsilon)h_1(y,t)\,,\\
\psi_2(y,t) &= \alpha(\alpha-\beta-\varepsilon)h_2(y,t)\,.
\end{aligned}$$

The marginal probability generating function for the process C_t^A becomes

$$C^A(y,t) = \frac{e^{\lambda t}\bar{K}_1(y)\phi_2(y,t) - \bar{K}_2(y)\phi_1(y,t)}{e^{\lambda t}\bar{K}_1(y)\psi_1(y,t) - \bar{K}_2(y)\psi_2(y,t)}\,, \tag{C.3}$$

where

$$\begin{aligned}
\bar{K}_1(y) &= -(\nu_1+\nu_0)(y-1)g_1(y,0)\,,\\
\bar{K}_2(y) &= (\alpha-\beta-\varepsilon)(\beta-\alpha)h_2(y,0) + (\beta\nu_1+\alpha\nu_0)(y-1)g_2(y,0)\,.
\end{aligned}$$

Case $\mu_\alpha = 0, \nu_1 = 0$

If we additionally have $\nu_1 = 0$ equation (4.17) becomes

$$\partial_s^2 Y - \frac{a_1 - 1}{s}\partial_s Y + \frac{a_2(a_0 + b_0 s)}{s^2} Y = 0\,.$$

Through the change of variables $z = 2\sqrt{a_2 b_0 s} = c\sqrt{s}$ and by seeking a solution of the form $Y(z) = z^{a_1} f(z)$ we reduce to the Bessel equation [3]

$$z^2 f'' + z f' + [z^2 - (a_2 - a_0)^2] f = 0\,. \tag{C.4}$$

The solution to equation (C.4) is given by the sum of two Bessel functions of first kind $J_\nu(z)$

$$f(z) = J_{a_2 - a_0}(z) + D J_{a_0 - a_2}(z)\,.$$

This allows us to write $Y(s)$ in terms of the functions $h_1(s) = J_{a_0 - a_2}(c\sqrt{s})$ and $h_2(s) = J_{a_2 - a_0}(c\sqrt{s})$ and since

$$\partial_z J_w(z) = \frac{w}{z} J_w(z) - J_{w+1}(z)\,,$$

we can also express the derivative of $Y(s)$ through h_1, h_2, g_1 and g_2, where

$$g_1(s) = J_{a_0 - a_2 + 1}(c\sqrt{s}), \quad g_2(s) = J_{a_2 - a_0 + 1}(c\sqrt{s})\,.$$

By joining such expression we first derive $X(s)$, and then find the integrating constant D by applying the initial condition $X(x, y, s = 1) = a_2 x$. Once again, we now multiply by $\frac{s}{a_2}$ and substitute $s = e^{-\varepsilon t}$ to find

$$\mathcal{A}(x, y, t) = \frac{K_1(x, y)\phi_2(y, t) - K_2(x, y)\phi_1(y, t)}{K_1(x, y)\psi_2(y, t) - K_2(x, y)\psi_1(y, t)}\,, \tag{C.5}$$

where

$$\begin{aligned}
K_1(x, y) &= (\alpha x - \alpha) h_1(y, 0) + \sqrt{\alpha \nu_0 (y - 1)}\, g_1(y, 0)\,,\\
K_2(x, y) &= (\alpha x - \beta) h_2(y, 0) + \sqrt{\alpha \nu_0 (y - 1)}\, g_2(y, 0)\,,\\
\phi_1(y, t) &= \alpha h_1(y, t) - \sqrt{\alpha \nu_0 (y - 1)}\, e^{-\frac{\varepsilon}{2} t} g_1(y, t)\,,\\
\phi_2(y, t) &= \beta h_2(y, t) - \sqrt{\alpha \nu_0 (y - 1)}\, e^{-\frac{\varepsilon}{2} t} g_2(y, t)\,,\\
\psi_1(y, t) &= \alpha h_1(t)\,,\\
\psi_2(y, t) &= \alpha h_2(t)\,.
\end{aligned}$$

From here we can recover the usual marginal probability generating function for the process A_t by considering the expansions of Bessel functions $J_w(z)$ around $z = 0$. The marginal probability generating function for the process C_t^A is instead given by

$$\mathcal{C}^A(y,t) = \frac{\bar{K}_1(y)\phi_2(y,t) - \bar{K}_2(y)\phi_1(y,t)}{\bar{K}_1(y)\psi_1(y,t) - \bar{K}_2(y)\psi_2(y,t)}\,, \tag{C.6}$$

where

$$\begin{aligned}\bar{K}_1(y) &= \sqrt{\alpha\nu_0(y-1)}g_1(y,0)\,,\\ \bar{K}_2(y) &= (\alpha-\beta)h_2(y,0) + \sqrt{\alpha\nu_0(y-1)}g_2(y,0)\,.\end{aligned}$$

Finally, we remark that in all the derivations above, there are special cases of the equation $f(z)$ that require extra attention. Namely, if some of the equation coefficients are integers, then its general solution assumes a slightly different expression than the one reported. However, it is unlikely that real estimates lead to such special cases, which therefore we do not discuss here.

C.2.2 Conditional distributions

In the summary of our model for biomarker shedding we pointed out that we are interested in the probability distributions of the processes A_t, B_t and C_t conditional on A_t non-extinction. Therefore, the asymptotic results derived in the previous sections are conditioned on the event Ω_∞^A. Here we show how a similar conditioning can be applied to the exact distributions.

Let us stress that all the probabilities involved in the following derivations are associated with the usual initial conditions for the processes A_t, B_t and C_t. For clarity we will not introduce a new notation for the corresponding conditioning, and hereafter we will thus implicitly denote

$$\mathbb{P}(\,\cdot\,) = \mathbb{P}\left(\,\cdot\;\middle|\;\left(A_0, B_0, C_0^A, C_0^B\right) = \left(1, B_0, 0, \frac{\nu_B}{\varepsilon}B_0\right)\right)\,.$$

Conditioning on A_t survival

As we are dealing with exact distributions at a given time, here we need to condition on the event of A_t survival up to time t, that is $\Omega_t^A = \{A_t > 0\}$. Let us denote $q_t^A = 1 - \mathbb{P}(\Omega_t^A) = \mathbb{P}(A_t = 0)$. Such a probability is equal to $\mathcal{A}(0,t)$,

and so we get (see also Section 1.2.1)

$$q_t^A = \frac{q_A - q_A e^{-\lambda_A t}}{1 - q_A e^{-\lambda_A t}} \,.$$

Bayes theorem yields

$$\mathbb{P}\left(A_t = m \mid \Omega_t^A\right) = \frac{\mathbb{P}\left(A_t = m\right)}{\mathbb{P}\left(\Omega_t^A\right)} \cdot \mathbb{P}\left(\Omega_t^A \mid A_t = m\right) \,,$$

for any $m \geq 0$. The term $\mathbb{P}\left(\Omega_t^A \mid A_t = m\right)$, however, is equal to 0 for $m = 0$ and to 1 for every $m \geq 1$. Hence, we get

$$\mathbb{P}\left(A_t = m \mid \Omega_t^A\right) = \frac{\mathbb{P}\left(A_t = m\right)}{1 - q_t^A} \,,$$

for every $m \geq 1$, and 0 otherwise. Similar steps allow to compute $\mathbb{P}\left(C_t^A = n \mid \Omega_t^A\right)$. In this case we find

$$\begin{aligned} \mathbb{P}\left(C_t^A = n \mid \Omega_t^A\right) &= \frac{\mathbb{P}\left(C_t^A = n, \Omega_t^A\right)}{\mathbb{P}\left(\Omega_t^A\right)} \\ &= \frac{\mathbb{P}\left(C_t^A = n\right) - \mathbb{P}\left(\left(A_t, C_t^A\right) = (0, n)\right)}{1 - q_t^A} \,. \end{aligned} \tag{C.7}$$

Here, the terms $\mathbb{P}(C_t^A = n)$ follow from the probability generating function $\mathcal{C}^A(y,t)$. Similarly, the probabilities $\mathbb{P}\left(\left(A_t, C_t^A\right) = (0,n)\right)$ can be obtained by inverting the function $\mathcal{A}(0,y,t)$.

Conditioning on one process size

The distribution of the total number of biomarker molecules present in the bloodstream at time t conditioned on the primary tumor size at that time is

$$\begin{aligned} \mathbb{P}(C_t = n \mid A_t = m) &= \sum_{i=0}^{n} \mathbb{P}\left(C_t^B = i\right) \mathbb{P}\left(C_t^A = n - i \mid A_t = m\right) \\ &= \frac{\sum_{i=0}^{n} \mathbb{P}\left(C_t^B = i\right) \mathbb{P}\left(A_t = m, C_t^A = n - i\right)}{\mathbb{P}(A_t = m)} \,. \end{aligned}$$

Once again, the terms $\mathbb{P}\left(A_t = m, C_t^A = n - i\right)$ follow from inverting the joint probability generating function $\mathcal{A}(x,y,t)$. Moreover in this case, if we consider a strictly positive primary tumor size m, conditioning on A_t survival is not necessary as $\{A_t = m\} \subset \Omega_t$. The exact probability mass function of C_t^A conditioned on the primary tumor size and its comparison with the asymptotic results obtained

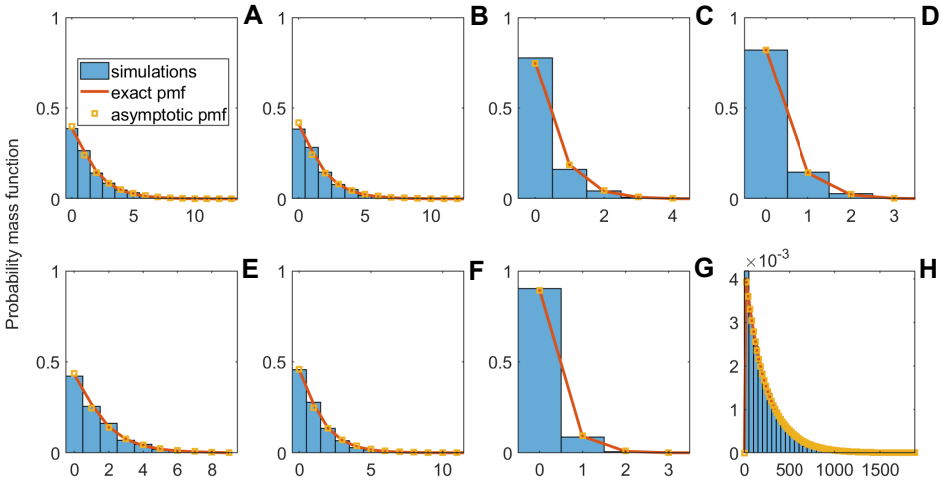

Figure C.1: ***Comparison between simulated biomarker levels, their exact and asymptotic theoretical distributions conditional on non-extinction of the primary tumor, with multiple shedding dynamics.*** *Panels (**A-G**) show the level of biomarker shed by cancer cells and still circulating in the bloodstream at time t for all the possible combinations of non-zero shedding rates. Bars illustrate the distribution of the number of biomarker molecules based on 10^4 simulations. The probability distributions illustrated are all conditional on primary tumor survival up to time t. Full red lines illustrate the exact probability distribution at time t derived from equations* (4.30), (C.3) *and* (C.6), *while the yellow squares represent the asymptotic results for large times and small shedding rate obtained from equation* (4.8). *Both perfectly agree with simulation results. Panel (**H**) similarly shows the probability distribution of the primary tumor size. Parameter values: birth rate $b = 0.051$ per cell per day and death rate $d = 0.041$ per cell per day; elimination rate $\varepsilon = 3.1 \times 10^{-5}$ per biomarker molecule per day; biomarker shedding probability per cell death $q_d = 10^{-4}$, per cell reproduction $q_b = 1.5 \times 10^{-4}$, and shedding rate per day (at cell necrosis) $\lambda_1 = 5 \times 10^{-5}$ per cell per day. All results are computed at time $t = 365$ days from the primary tumor onset.*

from equation (4.8) are illustrated in Figure C.1.

Now, once the exact probability mass function of C_t is known one can conversely condition on the total number of biomarker molecules present and ask how this affects the primary tumor size distributions. To see this we can write

$$\begin{aligned}\mathbb{P}(A_t = m \mid C_t = n) &= \frac{\mathbb{P}(A_t = m, C_t = n)}{\mathbb{P}(C_t = n)} \\ &= \frac{\sum_{i=0}^{n} \mathbb{P}\left(C_t^B = i\right) \mathbb{P}\left(A_t = m, C_t^A = n - i\right)}{\sum_{i=0}^{n} \mathbb{P}\left(C_t^B = i\right) \mathbb{P}\left(C_t^A = n - i\right)}.\end{aligned}$$

By further conditioning on A_t survival up to t we get

$$\begin{aligned}\mathbb{P}(A_t = m \mid C_t = n, \Omega_t^A) &= \frac{\mathbb{P}(A_t = m, \Omega_t^A \mid C_t = n)}{\mathbb{P}(\Omega_t^A \mid C_t = n)} \\ &= \frac{(1 - q_t^A)\mathbb{P}(A_t = m \mid C_t = n)}{\mathbb{P}(C_t = n \mid \Omega_t^A)\mathbb{P}(C_t = n)}\,.\end{aligned}$$

Finally, let us recall that as (B_t, C_t^B) is independent of the primary tumor growth dynamics, any kind of conditioning on A_t size has no effect on B_t and C_t^B probability distributions.

C.2.3 Expected values

The expected value and variance of the processes A_t, B_t and C_t at any given time t can immediately be computed by the corresponding generating functions (see Section 1.2.1). In the following we summarize formulas for the expectations, which are all implicitly conditional on $(A_0, C_0^A) = (1, 0)$ and $C_0^B = \frac{\nu_B}{\varepsilon} B_0$. For the cancerous cells population we have

$$\mathbb{E}[A_t] = e^{\lambda_A t}, \qquad \mathrm{Var}(A_t) = \frac{1 + q_A}{1 - q_A}\, e^{\lambda_A t}(e^{\lambda_A t} - 1)\,,$$

and

$$\mathbb{E}[A_t \mid \Omega_t^A] = \frac{e^{\lambda_A t} - q_A}{1 - q_A}, \quad \mathrm{Var}(A_t \mid \Omega_t^A) = \frac{(1 - q_A e^{-\lambda_A t})(1 + 2q_A)e^{\lambda_A t}(e^{\lambda_A t} - 1)}{(1 - q_A)^2}\,,$$

When the healthy population is modelled by a critical branching birth-death process we find

$$\mathbb{E}[B_t] \equiv B_0, \qquad \mathrm{Var}(B_t) = 2\beta_B B_0 t\,.$$

In this case, conditioning on A_t survival has no effect as A_t and B_t are independent processes. As for the number of biomarker molecules, from the functions $C^A(y,t)$ and $C^B(y,t)$ we get

$$\mathbb{E}[C_t^A] = \frac{\nu_A \left(e^{\lambda_A t} - e^{-\varepsilon t}\right)}{\lambda_A + \varepsilon}, \qquad \mathbb{E}[C_t^B] \equiv \frac{\nu_B}{\varepsilon} B_0\,.$$

In general, the expression for the latter mean would be the same for B_t modelled as a constant population or as a critical branching birth-death process. Furthermore, in our setup such expectation does not depend on t as we start the process C_t^B

at its equilibrium. Joining the previous results we find

$$\mathbb{E}[C_t] = \frac{\nu_A \left(e^{\lambda_A t} - e^{-\varepsilon t}\right)}{\lambda_A + \varepsilon} + \frac{\nu_B}{\varepsilon} B_0 \,.$$

The variance of C_t, and expressions for its first moments conditional on A_t survival can be obtained as well from equation (C.7), but they are too cumbersome to be reported here.

C.2.4 Sampling scheme

Here we derive the exact probability distribution for the number of biomarker molecules present in a blood sample of a given volume at time t. To this end, let us first assume that at t there is a fixed number $C_t = n$ of biomarker molecules uniformly distributed over a total volume V_{tot} of plasma. If we sample from it a volume V_s, each molecule is in the sample independently of the others with probability $p = \frac{V_s}{V_{tot}}$. Hence, the total number X_t of biomarker molecules present in the sample is binomially distributed with parameters n and p

$$\mathbb{P}(X_t = k \mid C_t = n) = \binom{n}{k} \left(\frac{V_s}{V_{tot}}\right)^k \left(1 - \frac{V_s}{V_{tot}}\right)^{n-k} \,. \tag{C.8}$$

The total number of biomarker molecules present in the plasma then follows by averaging over all the possible values of C_t

$$\mathbb{P}(X_t = k) = \sum_{n=k}^{\infty} \mathbb{P}(C_t = n)\mathbb{P}(X_t = k \mid C_t = n) \,. \tag{C.9}$$

The second term in the sum above is simply given by equation (C.8). As we noticed in the derivation of our asymptotic results, if C_t follows a Poisson distribution, then thanks to the thinning property X_t follows a Poisson distribution as well. The expected value and variance of X_t in terms of C_t are given by

$$\mathbb{E}[X_t] = p\,\mathbb{E}[C_t] \,, \quad \mathrm{Var}(X_t) = p(1-p)\mathbb{E}[C_t] + p^2\,\mathrm{Var}(C_t) \,.$$

The exact distribution of X_t can also be conditioned on A_t survival. First we have

$$\mathbb{P}\left(X_t = k \mid \Omega_t^A\right) = \frac{\mathbb{P}(X_t = k) - \mathbb{P}(X_t = k, A_t = 0)}{\mathbb{P}(\Omega_t^A)} \,. \tag{C.10}$$

Here, of the two terms at the numerator, the first one coincides with equation (C.9), while the second one can be expanded as

$$\begin{aligned}\mathbb{P}(X_t = k, A_t = 0) &= \sum_{n=k}^{\infty} \mathbb{P}(X_t = k, A_t = 0 \mid C_t = n)\mathbb{P}(C_t = n) \\ &= \sum_{n=k}^{\infty} \mathbb{P}(X_t = k \mid C_t = n)\mathbb{P}(A_t = 0 \mid C_t = n)\mathbb{P}(C_t = n)\,.\end{aligned}$$

The second equality follows from the fact that X and A_t are conditionally independent given C_t. Equation (C.10) thus becomes

$$\begin{aligned}\mathbb{P}(X_t = k \mid \Omega_t) &= \frac{\sum_{n=k}^{\infty} \mathbb{P}(X_t = k \mid C_t = n)\left[\mathbb{P}(C_t = n) - \mathbb{P}(A_t = 0, C_t = n)\right]}{\mathbb{P}(\Omega_t 6a)} \\ &= \frac{\sum_{n=k}^{\infty} \mathbb{P}(X_t = k \mid C_t = n)\mathbb{P}(C_t = n, \Omega_t^A)}{\mathbb{P}(\Omega_t^A)} \\ &= \sum_{n=k}^{\infty} \mathbb{P}(X_t = k \mid C_t = n)\mathbb{P}(C_t = n \mid \Omega_t^A)\,.\end{aligned}$$

Bibliography

[1] J. Abate and W. Whitt. Numerical inversion of probability generating functions. *Operations Research Letters*, 12(4):245–251, 1992.

[2] C. Abbosh, N. J. Birkbak, G. A. Wilson et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, 545(7655):446–451, 2017.

[3] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Applied mathematics series. Dover Publications, 1964.

[4] R. Adam, A. de Gramont, J. Figueras et al. Managing synchronous liver metastases from colorectal cancer: A multidisciplinary international consensus. *Cancer Treatment Reviews*, 41(9):729–741, 2015.

[5] S. Adeyemi and M. O. Ojo. A generalization of the Gumbel distribution. *Kragujevac J. Math*, 25:19–29, 2003.

[6] M. Ahsanullah, V. Nevzorov and M. Shakil. *An introduction to order statistics*, vol. 3. Atlantis Press, 1 edn., 2013.

[7] K. al Kattan, E. Sepsas, S. W. Fountain and E. R. Townsend. Disease recurrence after resection for stage I lung cancer. *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery*, 12:380–384, 1997.

[8] L. Allen. *An introduction to stochastic processes with applications to biology*. Boca Raton, FL: Chapman & Hall/CRC, 2 edn., 2011.

[9] P. D. Allison. *Survival Analysis Using SAS: A Practical Guide, Second Edition*. SAS Publishing, 2nd edn., 2010.

[10] P. L. Almeida and B. J. Pereira. Local treatment of metastatic prostate cancer: what is the evidence so far? *Prostate Cancer*, 2018:1–7, 2018.

[11] P. M. Altrock, L. L. Liu and F. Michor. The mathematics of cancer: integrating quantitative models. *Nat Rev Cancer*, 15(12):730–745, 2015.

[12] F. Andre, K. Slimane, T. Bachelot et al. Breast Cancer With Synchronous Metastases: Trends in Survival During a 14-Year Period. *Journal of Clinical Oncology*, 22(16):3302–3308, 2004.

[13] T. Antal and P. L. Krapivsky. Exact solution of a two-type branching process: clone size distribution in cell division kinetics. *J. Stat. Mech. Theory Exp.*, 2010(7):P07028, 22, 2010.

[14] T. Antal and P. L. Krapivsky. Exact solution of a two-type branching process: Models of tumor progression. *Journal of Statistical Mechanics P08018*, 2011.

[15] T. Arai, T. Kuroishi, Y. Saito et al. Tumor Doubling Time and Prognosis in Lung Cancer Patients: Evaluation from Chest Films and Clinical Follow-up Study. *Japanese Journal of Clinical Oncology*, 1994.

[16] R. P. Araujo and D. L. S. McElwain. A history of the study of solid tumour growth: the contribution of mathematical modelling. *Bulletin of Mathematical Biology*, 66(5):1039–1091, 2004.

[17] P. Armitage and R. Doll. The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis, 1954.

[18] P. Armitage and R. Doll. Stochastic Models for Carcinogenesis. Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp. 19–38. University of California Press, Berkeley, Calif., 1961.

[19] R. Askey and J. Wimp. Associated Laguerre and Hermite polynomials. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 96(1-2):15–37, 1984.

[20] K. B. Athreya and P. Ney. *Branching Processes*. Dover Publications, 2004.

[21] H. Awwad. *Radiation Oncology: Radiobiological and Physiological Perspectives*. Springer Netherlands, 2013.

[22] T. Bando. A new method of segmental resection for primary lung cancer: intermediate results. *European Journal of Cardio-Thoracic Surgery*, 21(5):894–899, 2002.

[23] N. Beerenwinkel, R. F. Schwarz, M. Gerstung and F. Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25, 2015.

[24] R. R. Berges, J. Vukanovic, J. I. Epstein et al. Implication of cell kinetic changes during the progression of human prostatic cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 1:473–480, 1995.

[25] F. Bidard, J. Madic, P. Mariani et al. Detection rate and prognostic value of circulating tumor cells and circulating tumor DNA in metastatic uveal melanoma. *International Journal of Cancer*, 134(5):1207–1213, 2014.

[26] D. Blackwell and D. Kendall. The Martin Boundary for Plya's Urn Scheme, and an Application to Stochastic Population Growth. *Journal of Applied Probability*, 1(2):284–296, 1964.

[27] S. Bolin, E. Nilsson and R. Sjdahl. Carcinoma of the colon and rectum - growth rate, 1983.

[28] A. Bolognese and L. Izzo. *Surgery in Multimodal Management of Solid Tumors*. Springer Milan, 2009.

[29] S. A. Boorjian, R. H. Thompson, M. K. Tollefson et al. Long-Term Risk of Clinical Progression After Biochemical Recurrence Following Radical Prostatectomy: The Impact of Time from Surgery to Recurrence. *European Urology*, 59(6):893–899, 2011.

[30] C. Boutros, C. Mazouni, F. Lerebours et al. A preoperative nomogram to predict the risk of synchronous distant metastases at diagnosis of primary breast cancer. *British Journal of Cancer*, 112(6):992–997, 2015.

[31] I. Bozic, T. Antal, H. Ohtsuki et al. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[32] I. Bozic, J. M. Gerold and M. A. Nowak. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLOS Computational Biology*, 12(2):e1004731, 2016.

[33] I. Bozic and M. A. Nowak. Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. *Proceedings of the National Academy of Sciences*, 111(45):15964–15968, 2014.

[34] F. Castro-Giner, S. Gkountela, C. Donato et al. Cancer Diagnosis Using a Liquid Biopsy: Challenges and Expectations. *Diagnostics*, 8(2):31, 2018.

[35] E. Çinlar. *Introduction to stochastic processes*. Prentice-Hall, 1975.

[36] C. L. Chaffer and R. A. Weinberg. A Perspective on Cancer Cell Metastasis. *Science*, 331(6024):1559–1564, 2011.

[37] D. Cheek and T. Antal. Mutation frequencies in a birth–death branching process. *The Annals of Applied Probability*, 28(6):3922–3947, 2018.

[38] R. Chignola and R. Foroni. Estimating the Growth Kinetics of Experimental Tumors From as Few as Two Determinations of Tumor Size: Implications for Clinical Oncology. *IEEE Transactions on Biomedical Engineering*, 52(5):808–815, 2005.

[39] S. J. Choi, H.-S. Kim, S.-J. Ahn, Y. M. Jeong and H.-Y. Choi. Evaluation of the growth pattern of carcinoma of colon and rectum by MDCT. *Acta Radiologica*, 54(5):487–492, 2013.

[40] T. Chou and M. R. D'Orsogna. First Passage Problems in Biology. In G. O. R. Metzler and S. Redner, eds., *First-Passage Phenomena and Their Applications*, pp. 306–345. World Scientific, 2014.

[41] G. L. Choudhury, D. M. Lucantoni and W. Whitt. Multidimensional Transform Inversion with Applications to the Transient M/G/1 Queue. *The Annals of Applied Probability*, 4(3):719–740, 1994.

[42] Z. Ciesielski and S. J. Taylor. First Passage times and Sojourn Times for Brownian Motion in Space and the Exact Hausdorff Measure of the Sample Path. *Transactions of the American Mathematical Society*, 103(3):434–450, 1962.

[43] V. P. Collins, R. K. Loeffler and H. Tivery. Observations on growth rates of human tumors. *The American journal of roentgenology, radium therapy, and nuclear medicine*, 76:988–1000, 1956.

[44] F. W. Crawford. *General birth-death processes: probabilities, inference, and applications*. Ph.D. thesis, Biomathematics, UCLA, 2012.

[45] D. J. Daley and D. Vere-Jones. *Introduction to the Theory of Point Processes*. Springer New York, 2006.

[46] A. V. D'Amico and G. E. Hanks. Linear regressive analysis using prostate-specific antigen doubling time for predicting tumor biology and clinical outcome in prostate cancer. *Cancer*, 72:2638–2643, 1993.

[47] H. David and H. Nagaraja. *Order Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, third edn., 2003.

[48] L. de Haan and A. Ferreira. *Extreme Value Theory, an introduction*. Springer Series in Operations Research and Financial Engineering. Springer, 2006.

[49] A. H. de l'Aulnoit, B. Rogoz, C. Pinçon and D. H. de l'Aulnoit. Metastasis-free interval in breast cancer patients: thirty-year trends and time dependency of prognostic factors. A retrospective analysis based on a single institution experience. *The Breast*, 37:80–88, 2018.

[50] F. C. Detterbeck and C. J. Gibson. Turning Gray: The Natural History of Lung Cancer Over Time. *Journal of Thoracic Oncology*, 3(7):781–792, 2008.

[51] P. Diaconis and L. Miclo. On Times to Quasi-stationarity for Birth and Death Processes. *Journal of Theoretical Probability*, 22(3):558–586, 2009.

[52] Z. Ding, Z. Wang, S. Huang, S. Zhong and J. Lin. Comparison of laparoscopic vs. open surgery for rectal cancer. *Molecular and Clinical Oncology*, 6(2):170–176, 2017.

[53] R. Durrett. *Branching process models of cancer*, vol. 1.1 of *Stochastics in biological systems*. Springer International Publishing, 1 edn., 2015.

[54] R. Durrett, J. Foo, K. Leder, J. Mayberry and F. Michor. Intratumor Heterogeneity in Evolutionary Models of Tumor Progression. *Genetics*, 188(2):461–477, 2011.

[55] A. Ebrahimi, J. R. Clark, N. Ahmadi et al. Prognostic significance of disease-free interval in head and neck cutaneous squamous cell carcinoma with nodal metastases. *Head & Neck*, 35(8):1138–1143, 2012.

[56] M. A. G. Elferink, K. P. de Jong, J. M. Klaase, E. J. Siemerink and J. H. W. de Wilt. Metachronous metastases from colorectal cancer: a population-based study in North-East Netherlands. *International Journal of Colorectal Disease*, 30(2):205–212, 2015.

[57] Y. E. Erdi. Limits of Tumor Detectability in Nuclear Medicine and PET. *Malecular Imaging and Radionuclide Therapy*, 21(2):23–28, 2012.

[58] R. Etzioni, N. Urban, S. Ramsey et al. The case for early detection. *Nature Reviews Cancer*, 3(4):243–252, 2003.

[59] M. Evans, N. Hastings and B. Peacock. *Statistical Distributions*. Wiley-Interscience, 3 edn., 2000.

[60] A. A. Farsi, A. Swaminath and P. Ellis. Patterns of Relapse in Small Cell Lung Cancer (SCLC): A Retrospective Analysis of Outcomes from a Single Canadian Center. *Journal of Thoracic Oncology*, 12(1):S727–S728, 2017.

[61] W. Feller. *An introduction to probability theory and its applications*, vol. 2. John Wiley & Sons, Inc, 1971.

[62] A. Ferlito, A. R. Shaha, C. E. Silver, A. Rinaldo and V. Mondin. Incidence and Sites of Distant Metastases from Head and Neck Cancer. *ORL*, 63(4):202–207, 2001.

[63] C. Fiala and E. P. Diamandis. Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection. *BMC Medicine*, 16(1):166, 2018.

[64] I. G. Finlay, D. Meek, F. Bruntont and C. S. McArdle. Growth rate of hepatic metastases in colorectal carcinoma. *British Journal of Surgery*, 75(7):641–644, 1988.

[65] D. J. Fitzpatrick, C. S. Lai, R. F. Parkyn et al. Time to Breast Cancer Relapse Predicted By Primary Tumour Characteristics, Not Lymph Node Involvement. *World Journal of Surgery*, 38(7):1668–1675, 2013.

[66] P. A. Fontenot Jr, A. Nehra, W. Parker et al. Metastatic prostate cancer in the modern era of PSA screening. *International Brazilian Journal of Urology*, 43(3):416–421, 2017.

[67] J. Foo and K. Leder. Dynamics of cancer recurrence. *The Annals of Applied Probability*, 23(4):1437–1468, 2013.

[68] J. F. Fowler. Biological Factors Influencing Optimum Fractionation in Radiation Therapy. *Acta Oncologica*, 40(6):712–717, 2001.

[69] S. Friberg and S. Mattson. On the growth rates of human malignant tumors: implications for medical decision making. *Journal of surgical oncology*, 65:284–297, 1997.

[70] S. Fujiwara, K. Yao, T. Nagahama et al. Can we accurately diagnose minute gastric cancers (5mm)? Chromoendoscopy (CE) vs magnifying endoscopy with narrow band imaging (M-NBI). *Gastric Cancer*, 18(3):590–596, 2015.

[71] D. Frnvik, K. Lång, I. Andersson et al. Estimates of breast cancer growth rate from mammograms and its relation to tumour characteristics. *Radiation Protection Dosimetry*, 169(1-4):151–157, 2015.

[72] E. Galante, G. Gallus, F. Chiesa et al. Growth rate of head and neck tumors. *European Journal of Cancer and Clinical Oncology*, 18(8):707–712, 1982.

[73] L. Gatteschi. Asymptotics and bounds for the zeros of Laguerre polynomials: a survey. *Journal of Computational and Applied Mathematics*, 144(12):7 – 27, 2002. Selected papers of the Int. Symp. on Applied Mathematics, August 2000, Dalian, China.

[74] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[75] G. Gundem, , P. V. Loo et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 2015.

[76] H. Haeno, M. Gonen, M. B. Davis et al. Computational Modeling of Pancreatic Cancer Reveals Kinetics of Metastasis Suggesting Optimum Treatment Strategies. *Cell*, 148(1-2):362–375, 2012.

[77] H. Haeno and F. Michor. The evolution of tumor metastases during clonal expansion. *Journal of Theoretical Biology*, 263(1):30–44, 2010.

[78] Y. C. Hagar, D. J. Harvey and L. A. Beckett. A multivariate cure model for left-censored and right-censored data with application to colorectal cancer screening patterns. *Statistics in medicine*, 35(PMC4938788):3347–3367, 2016.

[79] L. Hanin and L. Pavlova. Optimal screening schedules for prevention of metastatic cancer. *Statistics in Medicine*, 32(2):206–219, 2012.

[80] L. Hanin, A. Tsodikov and A. Yakovlev. Optimal schedules of cancer surveillance and tumor size at detection. *Mathematical and Computer Modelling*, 33(12-13):1419–1430, 2001.

[81] K. L. Harper, M. S. Sosa, D. Entenberg et al. Mechanism of early dissemination and metastasis in Her2+ mammary cancer. *Nature*, 540(7634):588–592, 2016.

[82] K. Haustermans, J. Fowler, K. Geboes et al. Relationship between potential doubling time (Tpot), labeling index and duration of DNA synthesis in 60 esophageal and 35 breast tumors: is it worthwhile to measure Tpot? *Radiotherapy and Oncology*, 46(2):157–167, 1998.

[83] K. M. Haustermans, I. Hofland, H. V. Poppel et al. Cell kinetic measurements in prostate cancer. *International Journal of Radiation Oncology∗Biology∗Physics*, 37(5):1067–1070, 1997.

[84] E. Heitzer, I. S. Haque, C. E. Roberts and M. R. Speicher. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics*, 2018.

[85] E. Heitzer, P. Ulz and J. B. Geigl. Circulating Tumor DNA as a Liquid Biopsy for Cancer. *Clinical Chemistry*, 61(1):112–123, 2014.

[86] C. I. Henschke, D. F. Yankelevitz, R. Yip et al. Lung Cancers Diagnosed at Annual CT Screening: Volume Doubling Times. *Radiology*, 263(2):578–583, 2012.

[87] P. Hohenberger, P. M. Schlag, T. Gerneth and C. Herfarth. Pre- and postoperative carcinoembryonic antigen determinations in hepatic resection for colorectal metastases. Predictive value and implications for adjuvant treatment based on multivariate analysis. *Annals of Surgery*, 219:135–143, 1994.

[88] J. W. Holch, M. Demmer, C. Lamersdorf et al. Pattern and Dynamics of Distant Metastases in Metastatic Colorectal Cancer, 2017.

[89] J. Hung, W. Jeng, W. Hsu et al. Prognostic factors of postrecurrence survival in completely resected stage I non-small cell lung cancer with distant metastasis. *Thorax*, 65(3):241–245, 2010.

[90] D. Hlzel, R. Eckel, R. T. Emeny and J. Engel. Distant metastases do not metastasize. *Cancer and Metastasis Reviews*, 29(4):737–750, 2010.

[91] M. E. Ismail, J. Letessier and G. Valent. Linear birth and death models and associated Laguerre and Meixner polynomials. *Journal of Approximation Theory*, 55(3):337 – 348, 1988.

[92] P. Jagers, F. C. Klebaner and S. Sagitov. Markovian paths to extinction. *Adv. in Appl. Probab.*, 39(2):569–587, 2007.

[93] K. S. Jain, A. G. Sikora, S. S. Baxi and L. G. T. Morris. Synchronous cancers in patients with head and neck cancer. *Cancer*, 119(10):1832–1837, 2013.

[94] A. R. Jensen, H. M. Nellemann and J. Overgaard. Tumor progression in waiting time for radiotherapy in head and neck cancer. *Radiotherapy and Oncology*, 84(1):5–10, 2007.

[95] N. L. Johnson, S. Kotz and N. Balakrishnan. *Continuous Univariate Distributions*, vol. 1 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons, Inc, second edn., 1994.

[96] S. B. Johnson, D. A. Hamstra, W. C. Jackson et al. Larger Maximum Tumor Diameter at Radical Prostatectomy Is Associated With Increased Biochemical Failure, Metastasis, and Death From Prostate Cancer After Salvage Radiation for Prostate Cancer. *International Journal of Radiation Oncology∗Biology∗Physics*, 87(2):275–281, 2013.

[97] S. Jones, W. Chen, G. Parmigiani et al. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences*, 105(11):4283–4288, 2008.

[98] S. Karlin and J. McGregor. The classification of birth and death processes. *Transactions of the American Mathematical Society*, 86(2):366–400, 1957.

[99] S. Karlin and J. McGregor. The differential equations of birth-and-death processes, and the stieltjes moment problem. *Transactions of the American Mathematical Society*, 85(2):489–546, 1957.

[100] S. Karlin and J. McGregor. Linear growth, birth and death processes. *J. Math. Mech.*, 7:643–662, 1958.

[101] S. Karlin and J. McGregor. Many server queuing processes with Poisson input and exponential service times. *Pacific Journal of Mathematics*, 8(1):87–118, 1958.

[102] S. Karlin and J. McGregor. Coincidence properties of birth and death processes. *Pacific Journal of Mathematics*, 9(4):1109–1140, 1959.

[103] S. Karlin and H. M. Taylor. *A first course in stochastic processes*. New York - San Francisco - London: Academic Press, Inc., a subsidiary of Harcourt Brace Jovanovich, Publishers. XVI, 1975.

[104] J. Keilson. Log-concavity and log-convexity in passage time densities of diffusion and birth-death processes. *Journal of Applied Probability*, 8(2):391–398, 1971.

[105] J. Keilson. *Markov Chain Models Rarity and Exponentiality*, vol. 28 of *Applied Mathematical Sciences*. Springer New York, 1979.

[106] P. Keller and T. Antal. Mutant number distribution in an exponentially growing population. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(1), 2015.

[107] M. M. Kemeny, S. Adak, B. Gray et al. Combined-Modality Treatment for Resectable Metastatic Colorectal Carcinoma to the Liver: Surgical Resection of Hepatic Metastases in Combination With Continuous Infusion of ChemotherapyAn Intergroup Study. *Journal of Clinical Oncology*, 20(6):1499–1505, 2002. PMID: 11896097.

[108] D. G. Kendall. Birth-and-death processes, and the theory of carcinogenesis. *Biometrika*, 47:13–21, 1960.

[109] K. M. Kerr and D. Lamb. Actual growth rate and tumour cell proliferation in human pulmonary neoplasms. *British Journal Of Cancer*, 50:343, 1984.

[110] D. A. Kessler and H. Levine. Scaling Solution in the Large Population Limit of the General Asymmetric Stochastic Luria–Delbrck Evolution Process. *Journal of Statistical Physics*, 158(4):783–805, 2014.

[111] H. Kim, D. H. Choi, W. Park et al. Prognostic factors for survivals from first relapse in breast cancer patients: analysis of deceased patients. *Radiation Oncology Journal*, 31(4):222, 2013.

[112] M. Kimmel and D. Axelrod. *Branching Processes in Biology*. Springer, New York, 2015.

[113] C. A. Klein. Parallel progression of primary tumours and metastases. *Nature Reviews Cancer*, 9:302, 2009.

[114] N. L. Komarova and D. Wodarz. Drug resistance in cancer: Principles of emergence and prevention. *Proceedings of the National Academy of Sciences*, 102(27):9714–9719, 2005.

[115] K. C. Koo, H. Yoo, K. H. Kim et al. Prognostic Impact of Synchronous Second Primary Malignancies on the Overall Survival of Patients with Metastatic Prostate Cancer. *Journal of Urology*, 193(4):1239–1244, 2015.

[116] P. Kornprat, M. J. Pollheimer, R. A. Lindtner et al. Value of Tumor Size as a Prognostic Variable in Colorectal Cancer. *American Journal of Clinical Oncology*, 34(1):43–49, 2011.

[117] S. Kotz and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. ICP, 2000.

[118] T. Kuroishi, S. Tominaga, T. Morimoto et al. Tumor Growth Rate and Prognosis of Breast Cancer Mainly Detected by Mass Screening. *Japanese Journal of Cancer Research*, 81(5):454–462, 1990.

[119] S. Kusama, J. S. Spratt Jr., W. L. Donegan, F. R. Watson and C. Cunningham. The gross rates of growth of human mammary carcinoma. *Cancer*, 30(2):594–599, 1972.

[120] D. Lea and C. Coulson. The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics*, 49(3):264–285, 1949.

[121] M. R. Leadbetter, G. Lindgren and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer New York, 1983.

[122] S. H. Lee, Y.-S. Kim, W. Han et al. Tumor growth rate of invasive breast cancers during wait times for surgery assessed by ultrasonography. *Medicine*, 95(37):e4874, 2016.

[123] S. P. Lee, J. Sun, H. Qian, W. H. McBride and H. R. Withers. Characterization of Metastatic Tumor Formation by the Colony Size Distribution. *arXiv pre-print*.

[124] H. J. Li, S. K. Ray, N. K. Singh, B. Johnston and A. B. Leiter. Basic helix-loop-helix transcription factors and enteroendocrine cell differentiation. *Diabetes, Obesity and Metabolism*, 13:5–12, 2011.

[125] E. Lianidou and K. Pantel. Liquid biopsies. *Genes, Chromosomes and Cancer*, 58(4):219–232, 2019.

[126] S. Liu, Y. Wong, J. Lin et al. Impact of recurrence interval on survival of oral cavity squamous cell carcinoma patients after local relapse. *Otolaryngology-Head and Neck Surgery*, 136(1):112–118, 2007.

[127] S. E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 48(6):491–511, 1943.

[128] P. M. Lykoudis, D. O'Reilly, K. Nastos and G. Fusai. Systematic review of surgical management of synchronous colorectal liver metastases. *Br J Surg*, 101(6):605–612, 2014.

[129] A. P. Makohon-Moore, K. Matsukuma, M. Zhang et al. Precancerous neoplastic cells can move through the pancreatic ductal system. *Nature*, 561(7722):201–205, 2018.

[130] K. Markou, J. Goudakos, S. Triaridis et al. The role of tumor size and patient's age as prognostic factors in laryngeal cancer. *Hippokratia*, 15(21607041):75–80, 2011.

[131] I. Martincorena and P. J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[132] Y. Masuda. First passage times of birth-death processes and simple random walks. *Stochastic Processes and their Applications*, 29(1):51 – 63, 1988.

[133] W. Q. Meeker and L. A. Escobar. *Statistical Methods for Reliability Data*. John Wiley & Sons Inc., 1998.

[134] R. Metzler, G. Oshanin and S. Redner. *First-passage phenomena and their applications*, vol. 35. World Scientific, 2014.

[135] F. Michor, M. A. Nowak and Y. Iwasa. Stochastic dynamics of metastasis formation. *Journal of Theoretical Biology*, 240(4):521–530, 2006.

[136] O. V. Motygin. On evaluation of the confluent Heun functions. In *2018 Days on Diffraction (DD)*. IEEE, 2018.

[137] A. Mustafin and E. Volkov. On the distribution of cell cycle generation times. *Biosystems*, 15(2):111–126, 1982.

[138] M. Muto, M. Nakane, C. Katada et al. Squamous cell carcinoma in situ at oropharyngeal and hypopharyngeal mucosal sites. *Cancer*, 101(6):1375–1381, 2004.

[139] H. N. Nagaraja. *Order Statistics from Independent Exponential Random Variables and the Sum of the Top Order Statistics*, pp. 173–185. Birkhäuser Boston, Boston, MA, 2006.

[140] K. Naxerova, E. Brachtel, J. J. Salk et al. Hypermutable DNA chronicles the evolution of human colon cancer. *Proceedings of the National Academy of Sciences*, 111(18):E1889–E1898, 2014.

[141] M. H. Neumann, S. Bender, T. Krahn and T. Schlange. ctDNA and CTCs in Liquid Biopsy – Current Status and Where We Need to Progress. *Computational and Structural Biotechnology Journal*, 16:190–195, 2018.

[142] M. D. Nicholson and T. Antal. Universal Asymptotic Clone Size Distribution for General Population Growth. *Bulletin of Mathematical Biology*, 78(11):2243–2276, 2016.

[143] B. Nordlinger, E. Van Cutsem, T. Gruenberger et al. Combination of surgery and chemotherapy and the role of targeted agents in the treatment of patients with colorectal liver metastases: recommendations from an expert panel. *Annals of Oncology*, 20(6):985–992, 2009.

[144] A. Novozhilov, G. Karev and E. Koonin. Biological applications of the theory of birth-and-death processes. *Brief Bioinform (2006)*, 7(1):70–85, 2006.

[145] T. Nowikiewicz, M. Wiśniewska, M. Wiśniewski et al. Overall survival and disease-free survival in breast cancer patients treated at the Oncology Centre in Bydgoszcz – analysis of more than six years of follow-up. *Współczesna Onkologia*, 4:284–289, 2015.

[146] A. C. Obenauf and J. Massagué. Surviving at a Distance: Organ-Specific Metastasis. *Trends in Cancer*, 1(1):76–91, 2015.

[147] A. G. Pakes. Asymptotic Results for the Extinction Time of Markov Branching Processes Allowing Emigration, I. Random Walk Decrements. *Advances in Applied Probability*, 21(2):243–269, 1989.

[148] A. G. Pakes. A Limit Theorem for the Maxima of the Para-Critical Simple Branching Process. *Advances in Applied Probability*, 30(3):740–756, 1998.

[149] J. H. Park, T. Kim, K. Lee et al. The beneficial effect of palliative resection in metastatic colorectal cancer. *British Journal Of Cancer*, 108:1425, 2013.

[150] C. A. Parkinson, D. Gale, A. M. Piskorz et al. Exploratory analysis of TP53 mutations in circulating tumour DNA as biomarkers of treatment response for patients with relapsed high-grade serous ovarian carcinoma: a retrospective study. *PLoS Medicine*, 13(12):e1002198, 2016.

[151] P. G. M. Peer, J. A. Van Dijck, A. L. Verbeek, J. H. Hendriks and R. Holland. Age-dependent growth rate of primary breast cancer. *Cancer*, 71(11):3547–3551, 1993.

[152] Y. Peng and J. M. Taylor. *Handbook of Survival Analysis*, chap. Cure Models, pp. 113–134. Chapman & Hall, 2014.

[153] S. Perakis and M. R. Speicher. Emerging concepts in liquid biopsies. *BMC Medicine*, 15(1):75, 2017.

[154] E. C. Pinheiro and S. L. P. Ferrari. A comparative review of generalizations of the Gumbel extreme value distribution with an application to wind speed data. *Journal of Statistical Computation and Simulation*, 86(11):2241–2261, 2016.

[155] N. F. Polizzi, M. J. Therien and D. N. Beratan. Mean First-Passage Times in Biology. *Israel Journal of Chemistry*, 56(9-10):816–824, 2016.

[156] L. Preziosi. *Cancer Modelling and Simulation*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press, Taylor & Francis Group, 2003.

[157] S. Redner. *A Guide to First-Passage Processes*. Cambridge University Press, 2001.

[158] J. G. Reiter, A. P. Makohon-Moore, J. M. Gerold et al. Minimal functional driver gene heterogeneity among untreated metastases. *Science*, 361(6406):1033–1037, 2018.

[159] A. A. Renshaw, J. P. Richie, K. R. Loughlin et al. Maximum diameter of prostatic carcinoma is a simple, inexpensive, and independent predictor of prostate-specific antigen failure in radical prostatectomy specimens. Validation in a cohort of 434 patients. *American journal of clinical pathology*, 111:641–644, 1999.

[160] E. Renshaw. *Modelling biological population in space and time*. Cambridge University Press, 1991.

[161] A. Rényi. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4(3):191–231, 1953.

[162] S. Resnick. *Extreme values, regular variation, and point processes*. Springer-Verlag, 1987.

[163] D. Rew and G. Wilson. Cell production rates in human tissues and tumours and their significance. Part II: clinical data. *European Journal of Surgical Oncology (EJSO)*, 26(4):405–417, 2000.

[164] A. Ronveaux. *Heun's Differential Equations*. Oxford University Press, 1995.

[165] S. Ross. *Stochastic processes*. John Wiley & Sons, Inc, 2 edn., 1996.

[166] S. M. Ross. *Introduction to Probability Models*. Elsevier LTD, Oxford, 2014.

[167] E. B. Ryu, J. M. Chang, M. Seo et al. Tumour volume doubling time of molecular breast cancer subtypes assessed by serial breast ultrasound. *European Radiology*, 24(9):2227–2235, 2014.

[168] E. Sahai. Illuminating the metastatic process. *Nature Reviews Cancer*, 7(10):737–749, 2007.

[169] J. D. Schiffman, P. G. Fisher and P. Gibbs. Early Detection of Cancer: Past, Present, and Future. *American Society of Clinical Oncology Educational Book*, 35:57–65, 2015.

[170] H. Schmid, J. E. McNeal and T. A. Stamey. Observations on the doubling time of prostate cancer. The use of serial prostate-specific antigen in patients with untreated disease as a measure of increasing cancer volume. *Cancer*, 71(6):2031–2040, 1993.

[171] M. Schwartz. A biomathematical approach to clinical tumor growth. *Cancer*, 14:1272–1294, 1961.

[172] J. G. Scott, P. Gerlee, D. Basanta et al. Mathematical Modeling of the Metastatic Process. In *Experimental Metastasis: Modeling and Analysis*, pp. 189–208. Springer Netherlands, 2013.

[173] S. Serres, M. S. Soto, A. Hamilton et al. Molecular MRI enables early and sensitive detection of brain metastases. *Proceedings of the National Academy of Sciences*, 109(17):6674–6679, 2012.

[174] R. Singh and K. Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, 2(4):145, 2011.

[175] I. N. Sneddon. On some infinite series involving the zeros of Bessel functions of the first kind. *Proceedings of the Glasgow Mathematical Association*, 4(3):144156, 1960.

[176] D. L. Snyder and M. I. Miller. *Random Point Processes in Time and Space*. Springer New York, 1991.

[177] J. S. Spratt and T. L. Spratt. Rates of Growth of Pulmonary Metastases and Host Survival. *Annals of Surgery*, 159(2):161–171, 1964.

[178] T. Strand. Survival after resection for primary lung cancer: a population based study of 3211 resected patients. *Thorax*, 61(8):710–715, 2006.

[179] C. Sturesson, V. T. Valdimarsson, E. Blomstrand et al. Liver-first strategy for synchronous colorectal liver metastases an intention-to-treat analysis. *HPB*, 19(1):52 – 58, 2017.

[180] U. Sumita. On Conditional Passage Time Structure of Birth-Death Processes. *Journal of Applied Probability*, 21(1):10–21, 1984.

[181] U. Sumita. On limiting behavior of ordinary and conditional first-passage times for a class of birth-death processes. *Journal of Applied Probability*, 24(1):235–240, 1987.

[182] R. M. Summers. Polyp Size Measurement at CT Colonography: What Do We Know and What Do We Need to Know? *Radiology*, 255(3):707–720, 2010.

[183] M. Tada, F. Misaki and K. Kawai. Growth rates of colorectal carcinoma and adenoma by roentgenologic follow-up observations. *Gastroenterologia Japonica*, 19:550–555, 1984.

[184] A. Talkington and R. Durrett. Estimating Tumor Growth Rates In Vivo. *Bulletin of Mathematical Biology*, 77(10):1934–1954, 2015.

[185] K. Tanaka, H. Shimada, M. Miura et al. Metastatic Tumor Doubling Time: Most Important Prehepatectomy Predictor of Survival and Nonrecurrence of Hepatic Colorectal Cancer Metastasis. *World Journal of Surgery*, 28(3):263–270, 2004.

[186] V. H. Teixeira, C. P. Pipinikas, A. Pennycuick et al. Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nature Medicine*, p. 1, 2019.

[187] Y. Tomimaru, S. Noura, M. Ohue et al. Metastatic Tumor Doubling Time Is an Independent Predictor of Intrapulmonary Recurrence after Pulmonary Resection of Solitary Pulmonary Metastasis from Colorectal Cancer. *Digestive Surgery*, 25(3):220–225, 2008.

[188] A. Toussi, S. B. Stewart-Merrill, S. A. Boorjian et al. Standardizing the Definition of Biochemical Recurrence after Radical Prostatectomy—What Prostate Specific Antigen Cut Point Best Predicts a Durable Increase and Subsequent Systemic Progression? *Journal of Urology*, 195(6):1754–1759, 2016.

[189] V. L. Tsikitis, D. W. Larson, M. Huebner, C. M. Lohse and P. A. Thompson. Predictors of recurrence free survival for patients with stage II and III colon cancer. *BMC Cancer*, 14(1), 2014.

[190] A. D. Tsodikov, J. G. Ibrahim and A. Y. Yakovlev. Estimating Cure Rates From Survival Data. *Journal of the American Statistical Association*, 98(464):1063–1078, 2003.

[191] M. Tnnies, J. Pfannschmidt, T. T. Bauer et al. Metastasectomy for Synchronous Solitary Non-Small Cell Lung Cancer Metastases. *The Annals of Thoracic Surgery*, 98(1):249–256, 2014.

[192] S. Umino, S. Hayashi and S. Ono. Doubling time of pulmonary metastases of adenoid cystic carcinoma. *International Journal of Oral and Maxillofacial Surgery*, 26:48, 1997.

[193] E. A. van Doorn. Representations for the rate of convergence of birth-death processes. Tech. Rep. 1584, 2001.

[194] E. A. van Doorn. Birth death processes and associated polynomials. *Journal of Computational and Applied Mathematics*, 153(12):497 – 506, 2003.

[195] E. Vinberg. *A course in algebra*, vol. 56 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[196] P. Vineis, A. Schatzkin and J. D. Potter. Models of carcinogenesis: an overview. *Carcinogenesis*, 31(10):1703–1709, 2010.

[197] D. von Fournier, E. Weber, W. Hoeffken et al. Growth rate of 147 mammary carcinomas. *Cancer*, 45:2198–2207, 1980.

[198] A. Waaijer, C. H. Terhaard, H. Dehnad et al. Waiting times for radiotherapy: consequences of volume increase for the TCP in oropharyngeal carcinoma. *Radiotherapy and Oncology*, 66(3):271–276, 2003.

[199] J. C. M. Wan, C. Massie, J. Garcia-Corbacho et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer*, 17:223, 2017.

[200] L. Wang. Early Diagnosis of Breast Cancer. *Sensors*, 17(7):1572, 2017.

[201] W. A. O. Waugh. Transformation of a Birth Process Into a Poisson Process. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(3):418–431, 1970.

[202] W. A. O. Waugh. The Apparent 'Lag Phase' in a Stochastic Population Model in which there is no Variation in the Conditions of Growth. *Biometrics*, 28(2):329–336, 1972.

[203] W. A. O. Waugh. Taboo Extinction, Sojourn Times, and Asymptotic Growth for the Markovian Birth and Death Process. *Journal of Applied Probability*, 9(3):486–506, 1972.

[204] W. A. O. Waugh. Uses of the sojourn time series for Markovian birth process. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Probability Theory*, pp. 501–514. University of California Press, Berkeley, CA, 1972.

[205] T. S. Weber, I. Jaehnert, C. Schichor, M. Or-Guil and J. Carneiro. Quantifying the Length and Variance of the Eukaryotic Cell Cycle Phases by a Stochastic Model and Dual Nucleoside Pulse Labelling. *PLoS Computational Biology*, 10(7):e1003616, 2014.

[206] P. N. Werahera, L. M. Glode, F. G. L. Rosa et al. Proliferative Tumor Doubling Times of Prostatic Carcinoma. *Prostate Cancer*, 2011:1–7, 2011.

[207] S. Wiegand, A. Zimmermann, T. Wilhelm and J. A. Werner. Survival After Distant Metastasis in Head and Neck Cancer. *Anticancer research*, 35:5499–5502, 2015.

[208] T. Williams. The Basic Birth-Death Model for Microbial Infections. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):338–360, 1965.

[209] M. S. Wilson, C. M. West, G. D. Wilson et al. Intra-tumoral heterogeneity of tumour potential doubling times (Tpot) in colorectal cancer. *British journal of cancer*, 68:501–506, 1993.

[210] S. Yachida, S. Jones, I. Bozic et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467:1114, 2010.

[211] A. Yakovlev. Threshold models of tumor recurrence. *Mathematical and Computer Modelling*, 23(6):153–164, 1996.

[212] A. Y. Yakovlev, A. D. Tsodikov and B. Asselain. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, 1996.

[213] U. Yilmaz and L. B. Marks. Estimating changes in the rate of synchronous and metachronous metastases over time: Analysis of SEER data. *Advances in Radiation Oncology*, 3(1):70–75, 2018.

[214] A. Yokoyama, N. Kakiuchi, T. Yoshizato et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*, p. 1, 2019.

[215] H. Yoo, B. Nam, H. Yang et al. Growth rates of metastatic brain tumors in nonsmall cell lung cancer. *Cancer*, 113(5):1043–1047, 2008.

[216] K. Zabicki, J. A. Colbert, F. J. Dominguez et al. Breast Cancer Diagnosis in Women $\leq$ 40 versus 50 to 60 Years: Increasing Size and Stage Disparity Compared With Older Women Over Time. *Annals of Surgical Oncology*, 13(8):1072–1077, 2006.

[217] B. Zackrisson, H. Gustafsson, R. Stenling, P. Flygare and G. D. Wilson. Predictive value of potential doubling time in head and neck cancer patients treated by conventional radiotherapy. *International Journal of Radiation Oncology∗Biology∗Physics*, 38(4):677–683, 1997.

[218] S. Zhang, Y. Ding, Q. Zhou et al. Correlation Factors Analysis of Breast Cancer Tumor Volume Doubling Time Measured by 3D-Ultrasound. *Medical Science Monitor*, 23:3147–3153, 2017.

[219] W. Zhang, W. Xia, Z. Lv et al. Liquid Biopsy for Cancer: Circulating Tumor Cells, Circulating Free DNA or Exosomes? *Cellular Physiology and Biochemistry*, 41(2):755–768, 2017.

[220] G. M. Zharinov, O. A. Bogomolov, N. N. Neklasova and V. N. Anisimov. Pretreatment prostate specific antigen doubling time as prognostic factor in prostate cancer patients. *Oncoscience*, 4:7–13, 2017.

[221] Q. Zheng. Progress of a half century in the study of the Luria–Delbrück distribution. *Mathematical Biosciences*, 162(1–2):1–32, 1999.