Journal of Mathematical Biology (2025) 90:65
https://doi.org/10.1007/s00285-025-02229-6

Mathematical Biology

A stochastic modelling framework for cancer patient trajectories: combining tumour growth, metastasis, and survival

Vincent Wieland[1,2] · Jan Hasenauer[1,2]

Received: 1 October 2024 / Revised: 16 March 2025 / Accepted: 30 April 2025 /
Published online: 22 May 2025

Abstract
Cancer is a major burden of disease around the globe and one of the leading causes of premature death. The key to improve patient outcomes in modern clinical cancer research is to gain insights into dynamics underlying cancer evolution in order to facilitate the search for effective therapies. However, most cancer data analysis tools are designed for controlled trials and cannot leverage routine clinical data, which are available in far greater quantities. In addition, many cancer models focus on single disease processes in isolation, disregarding interaction. This work proposes a unified stochastic modelling framework for cancer progression that combines (stochastic) processes for tumour growth, metastatic seeding, and patient survival to provide a comprehensive understanding of cancer progression. In addition, our models aim to use non-equidistantly sampled data collected in clinical routine to analyse the whole patient trajectory over the course of the disease. The model formulation features closed-form expressions of the likelihood functions for parameter inference from clinical data. The efficacy of our model approach is demonstrated through a simulation study involving four exemplary models, which utilise both analytic and numerical likelihoods. The results of the simulation studies demonstrate the accuracy and computational efficiency of the analytic likelihood formulations. We found that estimation can retrieve the correct model parameters and reveal the underlying data dynamics, and that this modelling framework is flexible in choosing the precise parameterisation. This work can serve as a foundation for the development of combined stochastic models for guiding personalized therapies in oncology.

Keywords Cancer modelling · Mathematical oncology · Stochastic modelling

Mathematics Subject Classification 92C50 · 92-10 · 62FXX · 62MXX

✉ Jan Hasenauer
jan.hasenauer@uni-bonn.de

[1] Life and Medical Science Institute, University of Bonn, Bonn, Germany

[2] Bonn Center for Mathematical Life Sciences, University of Bonn, Bonn, Germany

1 Introduction

Millions of people are diagnosed with cancer each year (World Health Organization 2018), making it the second most common cause of death worldwide. For the successful invention of new prevention measures and improvement of treatments, it is crucial to deepen our understanding of the dynamics underpinning the progression of cancer on the molecular as well as the macroscopic level (Elmore et al. 2021). Research solely based on clinical trials and traditional drug development pipelines faces the issue of high costs and ethical and regulatory roadblocks (Cui et al. 2020). Consequently, it is of importance to leverage the large amount of longitudinal data that is routinely recorded over the time course of their disease. This ranges from the initial diagnosis to data collected during monitoring of the patients throughout the therapy and disease progression. Such data can include patient characteristics and pre-diagnostic diseases and interventions, bloodcount data, or histological and radiological reports. However, data collected in clinical routine is collected at random timepoints and often incomplete and unstructured. Therefore, it is of paramount importance to construct mathematical and computational methods that can use this data with the objective of facilitating the search for an efficacious cure personalized for each patient (Rahman et al. 2022).

First groundbreaking attempts to aid this were made in the work of Collins et al. (1956) and Schwartz (1961) concentrating on modelling the tumour growth as the main driving force of cancer. Soon, other mathematical models of differing complexity followed (Laird 1964; Steel 1977). Comprehensive reviews on the plethora of different tumour growth models can be found in Gerlee (2013), Talkington and Durrett (2015), and Tjørve and Tjørve (2017).

A realistic model of cancer progression does not only require the description of the size of the tumour, but also various other hallmarks (Hanahan and Weinberg 2000; Hanahan 2011) such as tumour vascularization, acquisition of mutations, and the spread of metastasis. Indeed, metastatic cancer is responsible for the majority of cancer-related deaths (Fares et al. 2020). Due to further advances, in computing and cancer biology, larger computational models are possible, for example models for the mutation of key genes in cells (Gerlee 2013) or metastatic seeding (Franssen et al. 2019; Nguyen Edalgo and Ford Versypt 2018).

Besides modelling the driving forces of cancer progression in the human body, tumour growth and metastatic seeding, it is important to monitor and evaluate the survival of cancer patients and model its dependence on different variables such as treatment decisions. Therefore, many countries and initiatives have established policies to raise cancer awareness and screening programs for early detection of cancers, leading to a large amount of time-to-event data for the history of cancer progression (Zhang et al. 2023; Brito Fernandes et al. 2024). Analysing this data with classical parametric survival models lacks the flexibility to capture the complex underlying dynamics of the cancer disease (Perera and Dwivedi 2020). Moreover, selecting an appropriate model can be challenging (Palmer et al. 2023). In addition, more than one event or even the complete event history may be of interest, e.g. intermediate disease states may lead to different screening intervals. For these reasons multi-state models gained more attention in cancer screening evaluation and modelling cancer progres-

sion through different disease states (Cheung et al. 2022; Uhry et al. 2010). Still, they do not provide an understanding of the evolution of cancer progression in the patient.

As cancer modelling and screening capacities expanded, the speed of cancer therapy development increased; from classical chemotherapy evolving after Second World War, to drugs targeting specific molecules and lately to therapies using monoclonal antibodies and immune checkpoint inhibitors for the treatment of advanced or metastatic tumours (Arruebo et al. 2011; Falzone et al. 2018). Nowadays, the development personalized therapies that match the patient's individual case is of main focus in the field of medical oncology (Falzone et al. 2018). Also agent-based model gained attention in simulating the treatment decisions that need to be taken on the individual patient level (Mustapha et al. 2016; Calvaresi et al. 2020).

However, all of the afore-mentioned modelling approaches (Uhry et al. 2010; Nguyen Edalgo and Ford Versypt 2018; Franssen et al. 2019; Perera and Dwivedi 2020; Calvaresi et al. 2020) mostly capture a single aspect of cancer patient trajectories. To efficiently guide the developments in modern oncology, mathematical models for the analysis of patient trajectories should not focus on tumour growth, metastatic seeding, or patient survival, but integrate all these processes to provide a holistic view on the disease progression in the patient. In addition, inter-individual differences of patients and the stochastic nature of the stochastic processes should be captured. To meet these requirements, mathematical models needs to integrate different types of governing equations to accommodate the nature of the individual processes.

Hence, research that probabilistically links tumour growth and metastatic seeding based on clinical routine data to develop a more holistic view on the patient gained attention (Heimann and Hellman 2000; Minn et al. 2007; Gasparini and Humphreys 2022). One approach is the use of continuous growth based models. However, many of these are not flexible enough to be extended to adapt to different cancer entities or new processes (Isheden and Humphreys 2019). Another modelling approach tackling the task of providing a complete view on cancer progression is the use of discrete or continuous time multi-state Markov models. While Markov models can be easily extended, they quickly grow in complexity, which makes them computationally demanding and not suitable for practitioners (Uhry et al. 2010; Isheden and Humphreys 2019). Joint models of longitudinal and time-to-event data simultaneously describe disease dynamics by a non-linear mixed-effect model and use a survival model for the patient health status, which depends on unobserved biomarker kinetics (Rizopoulos 2011; Wu et al. 2012; Desmée et al. 2017). While they provide a good framework for individual dynamic predictions based on patient's covariates (Proust-Lima and Taylor 2009), they do not explicitly take into account the dependence of the different processes underlying the patient's disease progression.

In this work, we propose a new type of combined stochastic model that addresses the challenge of being efficient and flexible while maintaining a holistic view. To fill the gap in high-level models that describe the patient's trajectory based on routine clinical data in the existing model space, we provide a description of cancer progression and patient trajectories by expressing tumour growth, metastatic seeding, and patient survival with potentially stochastic processes. These three processes are then combined into a comprehensive model. The modelling framework is able to adapt to different assumptions about the underlying dynamics by exchanging the mathematical

description of the underlying processes. In addition, the proposed modelling framework does not rely on a specific kind of data, but facilitates the utilization of the rich source of information about cancer patients collected in the clinical routine and keeps the flexibility to incorporate patient specific effects in each sub-process separately. To ensure efficiency in the inference of the model parameters from the data, likelihood functions are calculated analytically, as far as possible.

This paper is organised as follows. In Sect. 2 we introduce our combined model which involves the three main processes of cancer progression and their mathematical representation together with examples for precise model formulations. Furthermore, we validate our model by showing that we can mimick characteristics of real-world data sets. Next, we introduce the likelihood function of the model for parameter inference from a given data set in Sect. 3. In Sect. 4, we showcase the inference capabilities of the proposed model type in a simulation study with the example formulations from Section 2. First, the accuracy and efficiency of the analytically computed likelihoods are assessed in Sects. 4.2 and 4.3. We then show that our estimation procedure retrieves the correct model parameters in Sect. 4.4. Section 4.5 displays the flexibility of the modelling framework by using more complex representations for the tumour growth. Furthermore, we illustrate the ability of the model to discover the true underlying dynamics of the data through model selection (Sect. 4.6) and to identify treatment and covariate effects (Sect. 4.7). Finally, we conclude the manuscript with a discussion and an outlook on further extensions to apply the framework in different real-world data settings.

2 Modelling framework

In order to provide mathematical models that can support and improve therapeutic decisions for patients, it is necessary to evaluate the progression of the cancer as well as the patient's health trajectory, i.e. health status. For this purpose, we propose a modelling framework that considers three dynamic processes: (1) the growth of a primary tumour, (2) the seeding of metastasis, and (3) the patient survival. These processes are time-continuous but may have a discrete or continuous state space, rendering analysis and inference challenging. To ensure the feasibility of statistical inference, we will focus on models with tractable likelihoods.

2.1 Modelling of cancer growth, metastatic spread and patient survival

In the following, we will first describe models for each of the processes separately (Sects. 2.1.1–2.1.2) and subsequently the combined model (Sect. 2.1.4).

2.1.1 Modelling tumour growth

To model the growth of the primary tumour, we rely on continuous growth models for tumour progression, which gained increasing attention as flexible alternatives to standard multi-state Markov models (Yin et al. 2019; Gasparini and Humphreys 2022;

Strandberg et al. 2023). These models express tumour size as a continuous function of time. Most commonly used are models for estimating the volume of the tumour, where tumour diameter measurements are recalculated into volume depending on assumptions of the spatial characteristics of the tumour (Talkington and Durrett 2015).

The first class of growth models we consider, consists of 1-dimensional ODEs modelling the total tumour size $S(t)$ by

$$\begin{aligned} \frac{dS}{dt} &= \beta S(t) f(t, S(t)), \\ S(0) &= S_0 \end{aligned} \tag{1}$$

in which β denotes the growth rate and the function f gives the deviation from a standard exponential growth (Talkington and Durrett 2015). The choice of f might introduce additional model parameters. The term $S(t) f(t, S(t))$ then takes into account the underlying biological suppositions about the growth behaviour of the tumour., e.g. constant cell-division, growth only occurring on the surface, or saturation of the tumour environment. To ensure that a solution of the ODE exists, we assume the following (Teschl 2012).

Assumption 1 The initial condition is positive, $S_0 > 0$, and $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous on $[S_0, \infty)$.

This general growth model (1) also includes the most commonly used formulations of tumour growth laws such as exponential growth, Gompertz growth, logistic growth, and power law representations (Gerlee 2013; Talkington and Durrett 2015; Spratt et al. 1993). In this work, we will use exponential and Gompertz growth models.

Exponential growth model The first model we use for modelling tumour growth is the simple exponential growth model, which resembles the assumption of a constant cell division rate independent of the tumour size. It was first applied to cancer in 1956 by Collins et al. (1956).

By choosing $f(t, S(t)) = 1$ in the general growth model (1), the exponential growth of a tumour is represented by the following differential equation

$$\frac{dS}{dt} = \beta S(t) \quad \text{with solution} \quad S(t) = S_0 e^{\beta t}. \tag{2}$$

where S_0 denotes the initial size of the tumour at detection t_0 and is fixed across all patients. This model works well in the context of small tumour sizes and limited time, where no external effects lead to a halt and saturation in particular does not yet occur (Plevritis et al. 2007; Abrahamsson and Humphreys 2016; Isheden and Humphreys 2019; Abrahamsson et al. 2020; Gasparini and Humphreys 2022; Ocaña-Tienda et al. 2024).

Gompertz growth model As an alternative to the simple exponential growth model, we consider a Gompertz growth model. This model was introduced in Gompertz (1825) to analyse human mortality curves. As a model for biologic growth, it resembles the decrease of the growth of an organism due to factors like saturation (Wright 1926).

Successfully applied to cancer by Laird (1964), it became commonly used, especially for breast cancer (Steel and Lamerton 1966; Steel 1977; Norton 1988, 2005; Tabassum et al. 2019).

We consider a Gompertz growth model with carrying capacity K expressed by the following ODE (Gerlee 2013).

$$\frac{dS}{dt} = \beta S(t) \ln\left(\frac{K}{S(t)}\right) \quad \text{with solution} \quad S(t) = K \exp\left(\ln\left(\frac{S_0}{K}\right)\exp(-\beta t)\right). \tag{3}$$

Gyllenberg–Webb model The modelling framework is not limited to 1-dimensional ODE models for tumour growth (1), but allows for more general continuous-time growth models. To demonstrate this, we consider in this work the physiologically structured tumour growth model by Gyllenberg and Webb (1990). This two-compartment model accounts for proliferating and quiescent cells, and is built on the assumptions that (1) actively proliferating cells can enter a quiescent state and (2) quiescence is more common in larger tumours. The model provides a generalization of Gompertz- and Bertalanffy-like S-shaped growth curves (Kozusko and Bajzer 2003; d'Onofrio et al. 2011; Kuang et al. 2018).

We consider the extended Gyllenberg–Webb model as proposed by (Alzahrani et al. 2014), which accounts for proliferating cells P, quiescent cells Q and dead cells R and is described by

$$\begin{aligned} \tfrac{dP}{dt} &= (b - r_0(S(t)))P(t) + r_i(S(t))Q(t), \\ \tfrac{dQ}{dt} &= r_0(S(t))P(t) - (r_i(S(t)) + \mu)Q(t), \\ \tfrac{dR}{dt} &= \mu Q(t) - dR(t), \\ M(t) &:= P(t) + Q(t) + R(t) \end{aligned} \tag{4}$$

where cells proliferate at per capita rate $b > 0$ and die at rates $\mu >= 0$. M gives the total tumour load. Additionally, cells transition between proliferating and quiescent compartment at rates $r_i(S), r_0(S)$ subject to some general assumptions (Gyllenberg and Webb 1990; Kuang et al. 2018). For this work we consider the example given in (Alzahrani and Kuang 2016)

$$r_i(M) = \frac{r}{M+m}, \quad r_0(M) = \frac{kM}{aM+1},$$

with $r = 1, k = 2, a = 1, m = 2$.

The Gyllenberg–Webb model is often used to include features related to resource limitation effects and yields a good starting point to explore treatment effect (Alzahrani et al. 2014; Kuang et al. 2018).

Since dead cells do not form metastases, we consider only proliferating and quiescent cells for the tumour volume $S(t)$. However, it is also possible to base the metastasis process only on one part of the tumour, e.g. only on proliferating cells, if biologically justified for the tumour entity at hand.

Example growth curves for the three tumour growth models are depicted in Fig. 1. A more exhaustive review on different forms of tumour growth laws with their respective

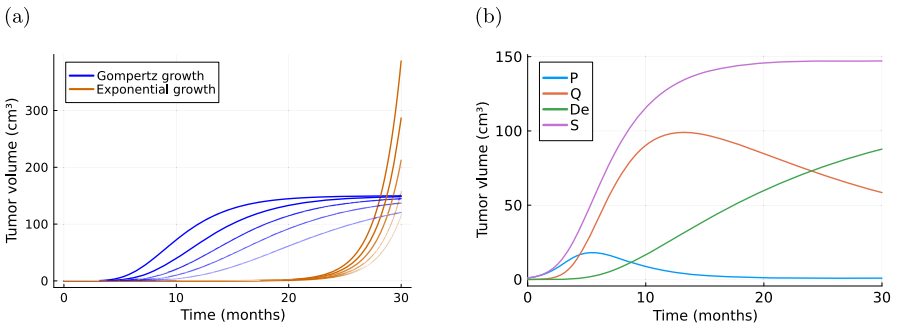


Fig. 1 Simulations of tumour growth curves. **a** Depicts example curves for exponential growth (2) and Gompertz growth (3). The initial tumour size is set to $S_0 = 0.065\,\text{mm}^3$ and saturation for the Gompertz growth to $K = 150\,\text{cm}^3$. For the corresponding growth parameters we chose $\beta \in \{0.48, 0.49, 0.5, 0.51, 0.52\}$ for exponential growth and $\beta \in \{0.14, 0.17, 0.2, 0.24, 0.3\}$ for the Gompertz growth indicated by the shade of the lines. **b** An example curve for the Gyllenberg–Webb model (4), where we choose $P(0) = 1$, $Q(0) = 0$, $De(0) = 0$ and example parameter values $[b = 1, \mu = 0.05, d = 0.01, r = 1, k = 2, a = 1, m = 2]$

advantages and interpretations can be found in (Marušić 1996; Gerlee 2013; Talkington and Durrett 2015; Kuang et al. 2018), and the references therein.

2.1.2 Modelling metastatic spread

Similar to the models of local spread to lymph nodes and metastatic seeding in the works of Gasparini and Humphreys (2022) and Isheden and Humphreys (2019), we exploit a stochastic model that models the number of metastases as $N(t) = N_0 + \hat{N}(t)$. Here, $\hat{N}(t)$ denotes an inhomogeneous Poisson point process. The rate of this Poisson process is chosen such that it accounts for the clinically established relationship between the size of the primary tumour and the spread of metastases (Koscielny et al. 1984; Sopik and Narod 2018). More precisely, we choose an intensity function $\lambda_N(t, S(t))$ which depends on both time and tumour size and N_0 denotes the number of metastases at detection time of the tumour. Thus the probability of n new metastasis seeding in a certain time-interval $[t_{j-1}, t_j)$ is given by

$$\mathbb{P}\left(N(t_j) - N(t_{j-1}) = n\right) = \frac{\Lambda_N([t_{j-1}, t_j), S)^n}{n!} \exp\left(-\Lambda_N([t_{j-1}, t_j), S)\right)$$

with

$$\Lambda_N([t_{j-1}, t_j), S) := \int_{t_{j-1}}^{t_j} \lambda_N(t, S(t))dt.$$

In order for this probability to be evaluable, we need the following assumption:

Assumption 2 The intensity function is integrable, i.e. $\lambda_N(t, S(t)) \in \mathrm{L}^1(\mathbb{R})$.

The rate of metastasis spread might depend on various factors, including number of cells, mutation stage, distance to vessel, and others. The intensity rate function λ_N can

be formulated in a way that explicitly takes these dependencies into account. Here, we consider two specific example rates focusing on the relation between the rate of metastatic seeding and the size of the primary tumour.

Volume based metastatic spread Metastatic seeding is an exceedingly complex process involving many different steps (Arvelo et al. 2016) and successful formation of metastasis depends on various not yet completely understood biochemical and genetic determinants that require research on their own (Hanahan 2011). As the number of cells which can leave the primary tumour to form metastasis depends on the size of the tumour, Bartoszynski et al. (2001) proposed as a first model the intensity $\lambda_N(t, S(t)) = bS(t)$. To account for the complexity of this process and other sources of metastatic seeding as well, we extend this and consider the following intensity

$$\lambda_N^{\text{vol}}(t, m_{\text{basal}}, m_{\text{size}}, S(t)) = m_{\text{basal}} + m_{\text{size}}\sqrt{S(t)}, \tag{5}$$

in which m_{basal} accounts for random metastatic seeding independent of tumour size and m_{size} is the coefficient of the effect of tumour size on the intensity. The dependence of metastatic seeding on observed characteristics can be incorporated by exchanging the constant parameters $m_{\text{basal}}, m_{\text{size}}$ by functions of various covariates such as age, medication, or others. The use of $\sqrt{S(t)}$ instead of $S(t)$ showed more numerically robust behavior in our simulation study in Sect. 4. However, it can be interchanged by other functions as well.

Cell division based metastatic spread One of the hallmarks of cancer is the capability of tumour cells to overcome hostile microenvironments when invading new systems (Hanahan 2011). Therefore, tumour cells need to be highly mutated to survive and establish a successful metastasis, which suggests a proportional relation of the rate of metastasis spread to the average number of mutations in the cancer cells and the rate of cancer cell division. Assuming a constant rate of mutation during cell division such a model has been proposed in Gasparini and Humphreys (2022) as an alternative to the volume based model intensity function (5). It is initially based on the model for lymph node spread by Isheden et al. (2019), where they describe the volume of a spherical, dense tumour without cell death through the number of cell divisions $A(t)$ and the size of a single cell S_{cell}

$$S_{\text{cell}} 2^{A(t)} = S(t)$$

yielding for the number of cell divisions $A(t)$ until a timepoint t

$$A(t) = \frac{\ln(S(t)) - \ln(S_{\text{cell}})}{\ln(2)}.$$

The rate of cell division in the tumour is then given by $A'(t)$. Using this, the intensity is denoted by

$$\lambda_N^{\text{div}}(t, m_{\text{division}}, S(t)) = m_{\text{division}} A(t)^k \frac{\partial A(t)}{\partial t} \tag{6}$$

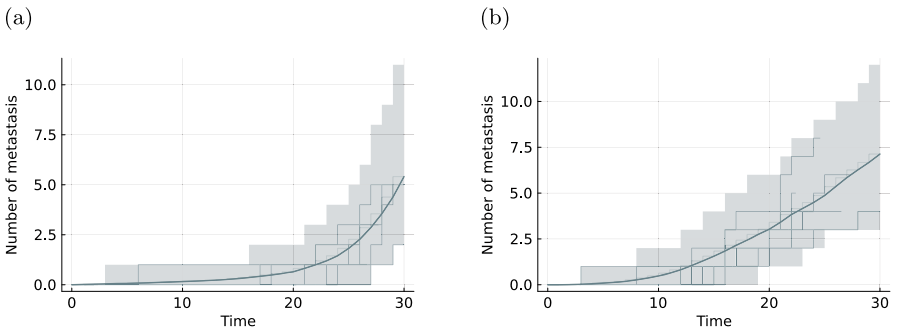


Fig. 2 Simulations of the metastasis process. Visualization of the mean metastasis number (solid line) with 95% pointwise credibility intervals (shaded regions) and 10 realisations (stepped lines) from 500 patients using the volume based metastatic spread (5) with exponential tumour growth (2) in **a** and using the Gompertz growth model (3) in **b**. The rates $m_{\text{basal}}, m_{\text{size}}$ of the metastasis intensity are the same in both cases

where the exponent $k \geq 1$ being added for additional flexibility.

We compared the effect of the two tumour growth models on the metastasis process by simulating two datasets consisting of 500 patients. For both datasets we used the same parameters of the metastasis process and the death process and calculated the mean metastasis number. The visual difference depicted in Fig. 2 showcases the influence of the choice of the tumour growth model on the process of metastatic seeding.

2.1.3 Modelling patient survival

The third process which we consider is the survival of the patient. This process can be influenced by several confounding factors such as age or general health status and the underlying mechanisms still remain unclear (Boire et al. 2024). However, the proximal causes of mortality in patients with cancer are often related to dysfunctional organs, for example, due to metastatic invasion (Fares et al. 2020). A plausible assumption to resemble this relationship in our model is that the survival rate depends on the growth of the tumour and the number of metastases. A more thorough analysis would require model-based approaches. Additionally, cancer patients are observed over a limited time frame and most likely die due to the disease progression and not general aging. Therefore, we will leverage the simplifying assumption of having exponentially distributed survival times. This means that the time a patient will survive does not depend on the time the patient already survived since diagnosis. Formally, this is named the "memoryless property" and translates directly into the death process being a Markov process.

Hence, we model the survival of the patient by considering the death process $D(t)$ to be an inhomogeneous Poisson point process that is stopped after the first jump occurs. The corresponding intensity rate is denoted by $\lambda_D(t, N(t), S(t))$. Thus the probability to survive until some time $t > 0$ is given by

$$\mathbb{P}(D(t) = 0) = \exp\left(-\Lambda_D(t, N(t), S(t))\right)$$

with

$$\Lambda_D(t, N(t), S(t)) := \int_0^t \lambda_D(s, N(s), S(s))ds.$$

The intensity rate function λ_D depends on the tumour size and the metastatic load in the patient. As before we will assume integrability of λ_D. We assume the following linear dependence on the tumour growth and metastasis number throughout the rest of this work

$$\lambda_D(t, d_{\text{size}}, d_{\text{metas}}, S(t), N(t)) = d_{\text{size}}\sqrt{S(t)} + d_{\text{metas}}N(t), \tag{7}$$

where d_{size} and d_{metas} quantify the influence of the corresponding process on the survival time.

2.1.4 Combined stochastic model for cancer growth, metastatic spread and associated death

As the three previously processes are essential to describe and understand cancer patient survival, we combine them into one model providing a holistic description of the cancer patient. The resulting combined stochastic process, $\mathbf{X}(t) = (S(t), N(t), D(t))^T$, provides a novel formulation of a mathematical model for cancer progression. The Markovian nature of all three subprocesses ensures that the combined stochastic process still satisfies the Markov property. Furthermore, all subprocesses admit dependencies with each other, since the last observation timepoint is determined by the death process. A graphical exemplification using the example of breast cancer modeled with exponential tumour growth (2) and volume-based metastasis intensity (5) can be seen in Fig. 3.

2.1.5 Individual-specific parameterisation

Cancer as a disease represents a complex ecosystem of hundreds of distinct types, which can vary substantially in their behavior. Even if one focuses on the analysis of one specific type of cancer, e.g. breast cancer, the progression varies between individual patients. In the above sections, we introduced rate parameters determining the dynamics of the process $\mathbf{X}(t)$ of interest, e.g. a tumour growth rate β. The combined stochastic model describes the stochastic evolution of cancer in a homogeneous population. Yet, it is well established that there are inter-individual differences which influence tumour growth rates β, metastasis development rates $m_{\text{basal}}, m_{\text{size}}, m_{\text{div}}$ and other properties. Therefore, we allow the rates of the processes to depend on individual-specific, potential time-dependent covariates $\mathbf{Z}$. The covariates for patient l are denoted by $\mathbf{Z}^l(t)$, and might encode treatment, age, and weight.

Modelling treatment effect on tumour growth Using the presented framework, we can investigate the effect of a treatment aimed at the growth of the primary tumour. We can model this by expressing the tumour growth rate β with a linear model of a

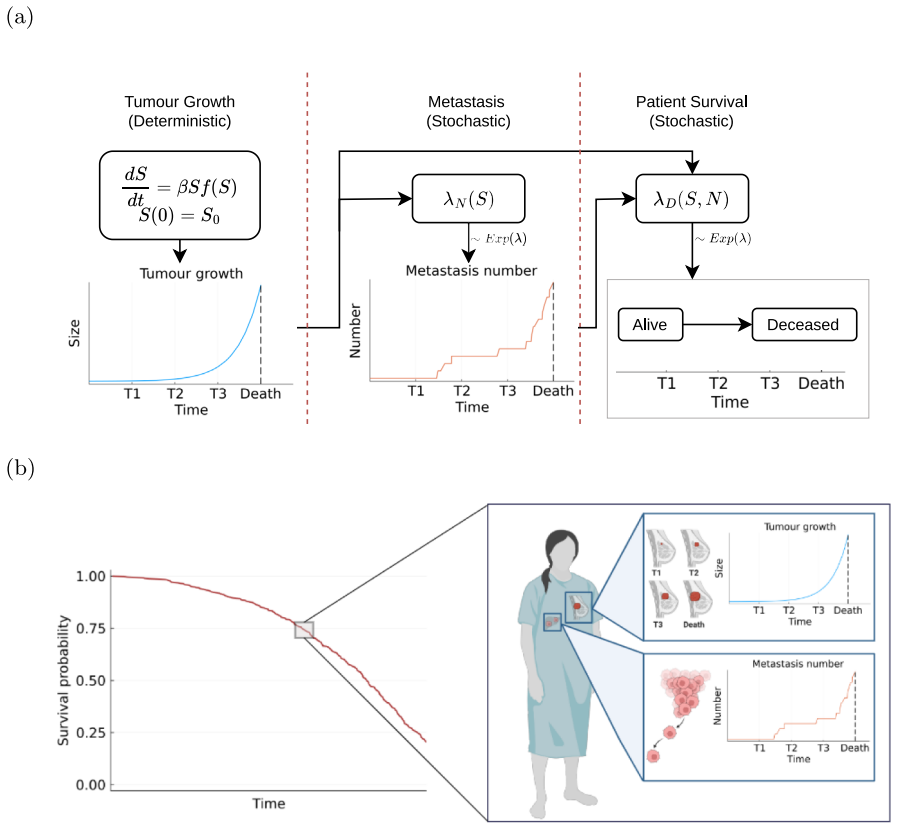


Fig. 3 Illustration of the proposed model of cancer patient trajectories. **a** Outline of the three components of the model, the tumour size S with growth parameter β, the metastasis process N with rate $\lambda_N(S)$ and the death process D with rate $\lambda_D(S, n)$ and their interconnections. **b** Simulation results for a population of 500 cancer patients using parameters reported in Supplementary Table S2: (left) Survival curve and (right) tumour growth and metastasis development for an exemplary patient. Created in BioRender. Hasenauer, AG. (2024) BioRender.com/k04j643.

basal growth rate and a covariate dependent effect. Both, the treatment status and other covariates that might influence the treatment effect are summarised in the covariate vector $\mathbf{Z}^l(t)$ and the rate parameter is then given by

$$\beta(t, \mu, \mathbf{Z}^l) = \beta_0 + \mu^T \mathbf{Z}^l(t), \tag{8}$$

where μ^T is the vector of effects of treatment and covariates on the tumour growth. This linear relationship of covariates and model parameters can also be replaced by other functional representations.

With all model choices made, the parameters of interest are β_0, μ^T. We collect those together with the parameters of the other two processes, metastasis spread and patients' survival, in the model parameter vector $\boldsymbol{\theta}$ and denote the disease trajectory of patient l by $\mathbf{X}(t, \boldsymbol{\theta}, \mathbf{Z}^l)$.

2.2 Observation model

The state of the cancer progression within a patient cannot be observed continuously and without measurement noise. Accordingly, the model for the disease dynamics needs to be complemented by a model for the observation process. Here, we consider the case that the observation time points are determined by the visit time point of the patient to the hospital. Thus, the time points for the assessment of the size of the primary tumour and the number of metastases are independent from the two processes.

In clinical practice, the tumour size is mostly inferred by the use of imaging techniques rather than pathological measurements. Different factors such as radiographic imaging resolution or physician contouring preferences might lead to noise-corrupted data of the tumour size $\tilde{S}(t)$ (Harshe et al. 2023; Jakubowski et al. 2012; Paquelet and Hendrick 2010). We represent this by adding a normally distributed random variable with size-dependent variance $\sigma_S^2 S(t)^2$ to the tumour size

$$\tilde{S}(t) = S(t) + \varepsilon, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma_S^2 S(t)^2). \tag{9}$$

This assumption of normally distributed independent measurement noise can also be exchanged with different noise model assumptions. The application of such models needs to account for the measurement techniques and has to capture its characteristics in the observation function and the noise distribution. For the purpose of this work, the only requirement on the noise model is that it admits an analytical representation of the corresponding log-density function. To showcase the flexibility of our framework in the choice of the noise model, we provide an alternative formulation using log-normally distributed tumour size measurements in the Supplementary Information A.1.

Since the observation time points are uninformative and independent of the process, we assume that we observe the number of metastases $\tilde{N}(t)$ at each screening timepoint. In clinical practice, metastases, just like the primary tumour, must reach a certain detection threshold in order to be observed. However, under the assumption of constant growth rates for the metastases, this would only be a constant time delay in the detectability of events and will not influence the time passing between two events. Therefore, we assume that the metastasis process jumps as soon a metastasis reaches this threshold and we get observations $\tilde{N}(t_i) = N(t_i)$ for the i-th observation time point.

As opposed to these observations at uninformative and process independent screening timepoints, the death is observed directly when it occurs. Hence, we observe the actual time of the transition of the death process from alive $D(t) = 0$ to dead $D(t) = 1$ and the last measurement timepoint always coincides with the timepoint of death T_d of the corresponding patient.

2.3 Implementation of the combined model

The code for the model implementation was written in the Julia programming language version 1.10.4 (Bezanson et al. 2017). For the formulation and simulation of the differential equation for the tumour growth, we used the `Differentialequations.jl`

package from the SciML ecosystem (Rackauckas and Nie 2017). On top of this, we implemented a version of the next reaction method based on Anderson (2007) to simulate the two inhomogeneous Poisson point processes depending on the solution of the differential equation.

2.4 Analysis of the combined model and validation on published data

In the previous sections, we introduced a novel model for the description of clinical trajectories. To assess the plausibility of this model, we compare (1) the assumption about the subprocesses with real-world data on patient trajectories, and (2) the properties of the overall model with real-world data from clinical trials.

Assessment of assumptions on subprocess' properties

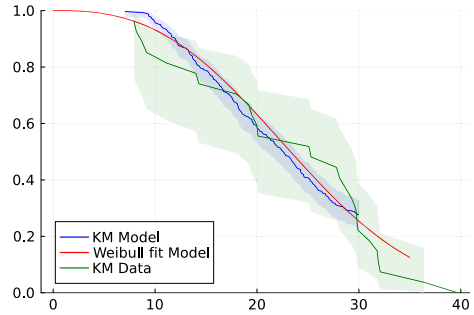
So far we introduced a modelling framework and presented simple models for the subprocesses. While these subprocess models can be easily replaced, we hypothesize that some assumptions might appear more restrictive than they actually are as the coupling within the combined model allows for additional degrees of freedom.

To assess this hypothesis, we consider the assumption of a simple Markov processes for survival. On first glance, this might suggest that the survival tile follows an exponential distributions, which has been shown to be unrealistic for many tumour types (Baghestani et al. 2016; Klakattawi 2022). Yet, as the simple survival process is coupled with a growth model for the tumour size and a metastasis process, the distribution of the survival times does not admit a simple exponential distribution as seen in the histograms of the overall survival times (Supplementary Figure S1). Since for each patient the overall survival time consists of a sequence of inter-transition times with survival rates being influenced by the tumour growth and the jump process of metastasis count, one faces a sequence of exponentially distributed times with differently time-varying rates. To investigate the distribution of the overall survival we fitted a general survival time distribution using a maximum-likelihood approach for censored data. Namely, we used the generalized Gamma distribution[1] to describe the survival times obtained by simulating two models since it inherits common survival time distributions such as the exponential distribution, the Weibull distribution and the Gamma distribution as special cases (Box-Steffensmeier and Jones 2004). The first model we used combined exponential growth with the proportional metastasis intensity and the second used Gompertz growth to describe the tumour size. The survival time distribution for the first model closely resembles the a generalized Gamma distribution (Supplementary Figure S2). The survival time distribution for the second model showed to be close to a Weibull distribution based on the one-sample Kolmogorov–Smirnov test (K-S test) and Pearsons chi-squared test (Agresti 2013). The Weibull distribution is commonly used in describing survival of cancer patients and treatment effects on it (Baghestani et al. 2016; Plana et al. 2022; Klakattawi 2022).

To study the capabilities of the model to describe real-world observational data, we consider the ACT2-trial (Hagman et al. 2016). In this trial, colorectal cancer patients with KRAS mutations were treated with ACT2. The dataset has been published Can-

[1] Parametrization given in the Supplementary Information A.2.

(a)



(b)

Parameter	Value
K	100000
α	0.15
m_{basal}	0.01
m_{size}	0.001
d_{size}	0.003
$d_{\text{metastasis}}$	0.003

Fig. 4 Assessment of survival times distribution. **a** Provides a visualization of the estimated survival curve of model (M4) combining Gompertz growth with exponential proportional metastasis intensity (blue) against the estimated survival curve of 33 patients from the ACT2-trial (green). Additionally, the survival curve of a Weibull distribution with parameters $k = 2.7$, $\lambda = 26.7$ (red) is plotted. **b** Provides the parameter values used to simulate the data from the model

certrials.io (2022) and was previously fitted with a Weibull distribution Plana et al. (2022). The Gompertz growth based model was able to mimick the survival curve of this trial as shown in Fig. 4. MLE-based distribution fitting showed that a Gompertz distribution with parameters $k = 2.7$, $\lambda = 26.7$ describes the survival data from the ACT2-trial and the survival times obtained by our model with the parameters given in Fig. 4bbased on the K–S and Pearson's chi-squared test.

Assessment of proposed model structure

To validate the proposed model structure in the context of real-world data, we investigate whether the model can recover characteristics of real-world datasets. For this purpose, we considered the work of Engel et al. (2003), in which they analyzed 12,423 patients with breast cancer from the Munich Cancer Registry (MCR)[2] from 1978 to 1996. They stratified the patients into the four T stages pT1, pT2, pT3, pT4 of the TNM classification system depending on the initial size of the primary tumour at time of diagnosis. As one of the results they reported that the overall survival following diagnosis varies significantly across these groups (Figure 1 in (Engel et al. 2003)). Whereas survival after metastasisation appears to be independent of the pT category, indicating an almost homogeneous growth of metastases (Figure 3 in (Engel et al. 2003)). The model combining exponential tumour growth with the cell division based metastasis rate (6) can reproduce these findings. We simulated 4 datasets of 1000 patients which only differed in the initial tumour size but not the model parameters. Initial tumour sizes were chosen to represent the four pT categories with corresponding diameters 10 mm, 35 mm, 75 mm and 120 mm and the model parameters are given in Supplementary Table S1. Only using different initial tumour sizes the model is capable of representing the finding that survival after metastasisation is nearly independent of the primary tumour size at time of diagnosis, but over survival differs significantly across the four groups. Additionally, we used the WebPlotDigitizer software (Ankit

[2] https://www.tumorregister-muenchen.de/en/.

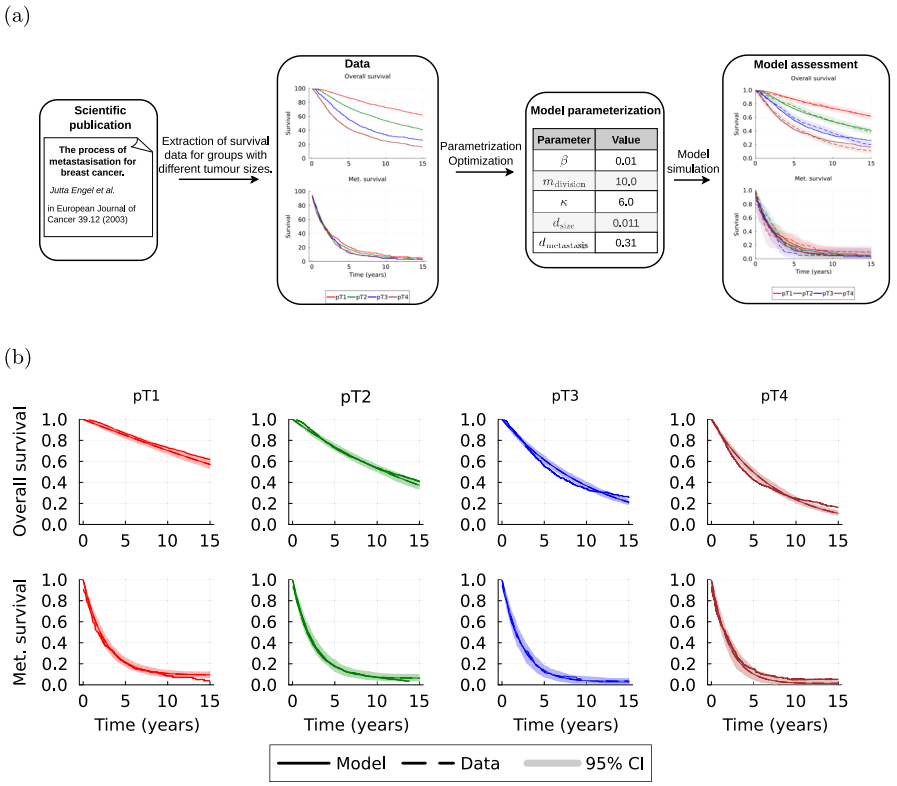


Fig. 5 Assessment of proposed model structure. **a** Visual outline of model assessment performed using the dataset by Engel et al. (2003) on overall survival and survival after metastasisation of breast cancer patients across different tumour sizes at time of diagnosis. **b** Comparison of observed (data) and simulated (model) survival curves. The simulation results are indicated using the mean survival curve (line) and the 95-percentile interval (shaded area) from 200 simulations of our model

2025) to recreate the survival curves for overall survival and survival after metastasisation from Figure 1 and Figure 3 in (Engel et al. 2003). Figure 5 shows that the survival curves obtained from the simulated datasets with our proposed model show a consistent behavior with the ones reported.

3 Parameter inference

The combined stochastic model for a cancer patient depends on several model parameters determining the tumour growth rate, metastatic seeding rate, and survival rate. To correctly assess disease progression, impact of treatment effects and related properties, it is necessary to infer these parameters characterising the evolution of a patient from their corresponding clinical data. This can be achieved using likelihood-based and likelihood-free approaches. Likelihood-free such as approximate Bayesian computation are computationally expensive (Schälte et al. 2022), leaving likelihood-based

approaches as the method of choice. Yet, they require a likelihood formulation which admits computationally efficient evaluation.

Here, we derive such formulations for the aforedescribed combined stochastic model: a fully analytical expression and an expression which admits efficient numerical integration. Subsequently, we describe how these formulations are used for parameter optimisation and uncertainty analysis.

3.1 Computation of the likelihood function

We consider the inference of the model parameters from the clinical data for a population of K patients, $\mathcal{D} = \{\mathcal{D}^l\}_{l=1}^K$. As the clinical datasets for individual patients are independent the likelihood can be factorized as follows

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \prod_{l=1}^{K} p(\mathcal{D}^l|\boldsymbol{\theta}^l).$$

In the following, we outline without loss of generality how the likelihood contribution from patient l, $p(\mathcal{D}^l|\boldsymbol{\theta}^l, \mathbf{Z}^l)$, can be evaluated. For ease of notation we omit the patient index in the rest of the section as well as the possible dependence on patient specific covariates available in the data.

For a single patient, the likelihood describes the probability of observing the corresponding dataset $\mathcal{D} = \{(\mathbf{y}_j, t_j)\}_{j=1}^N$, consisting of observations $\mathbf{y}_j = (\tilde{S}_j, \tilde{N}_j, \tilde{D}_j)^T$ at timepoints $\{t_j\}_{j=1}^N$, given a parameter vector $\boldsymbol{\theta}$,

$$p(\mathcal{D}|\boldsymbol{\theta}) = \int_{\Omega_\theta} p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{X})\, p_{\boldsymbol{\theta}}(\mathbf{X}) d\mathbf{X}, \tag{10}$$

where Ω_θ is the full state-space of the combined stochastic process $\mathbf{X}$ for a given parameter vector $\boldsymbol{\theta}$. The first term in the integral is the likelihood of the observation model and the second term the likelihood contribution of the hidden process.

Due to the Markov property of $\mathbf{X}$, the observation timepoints divide the sample path of the process into mutually independent parts. Therefore, we obtain the likelihood for one patient by factorising over the observation timepoints.

$$\int_{\Omega_\theta} p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{X})\, p_{\boldsymbol{\theta}}(\mathbf{X}) d\mathbf{X} = \int_{\Omega_\theta} \prod_{j=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{y}_j|\mathbf{X}(t_j))\, p_{\boldsymbol{\theta}}(\mathbf{X}(t_j)|\mathbf{X}_{t_{j-1}}) d\mathbf{X}. \tag{11}$$

Likelihood of the observation model

The first term in the product on the right-hand side of (11), $p_{\boldsymbol{\theta}}(\mathbf{y}_j|\mathbf{X}(t_j)) = p(\mathbf{y}_j|\mathbf{X}(t_j), \boldsymbol{\theta})$, is the probability of observing $\mathbf{y}_j$ given the actual state $\mathbf{X}(t_j) = (S(t_j), N(t_j), D(t_j))^T$ of the process at observation timepoint t_j. Due to the assumptions made in Sect. 2.2, we exactly observe the number of current metastases and

the death timepoint T_d of the patient. The latter is equivalent to observing the current survival status of the patient $D(t_j)$. This yields $\mathbb{P}(\tilde{N}_j|N(t_j)) = \delta_{\tilde{N}_j,N(t_j)}$ and $\mathbb{P}(\tilde{D}_j|D(t_j)) = (1-\delta_{t_j,T_d})$, where $\delta_{i,j}$ denotes the Kronecker delta. Hence, the only contribution to the likelihood of the observation process is given by the measurement noise in the tumour size measurements. Therefore, we can write the likelihood contribution of the observation model as

$$p_{\boldsymbol{\theta}}(\mathbf{y}_j|\mathbf{X}(t_j)) = \delta_{\tilde{N}_j,N(t_j)}(1-\delta_{t_j,T_d})\frac{1}{\sqrt{2\pi(\sigma_S\cdot S(t_j))^2}}\exp\left(-\frac{(S(t_j)-\tilde{S}_j)^2}{2(\sigma_S\cdot S(t_j))^2}\right).$$

This yields that the integral in (11) vanishes on every possible path of $\mathbf{X}$ except for the set of paths exactly matching the observations of metastasis process and survival which we denote by $\tilde{\Omega}_{\boldsymbol{\theta}} = \{\mathbf{X}\in\Omega_{\boldsymbol{\theta}}|N(t_j)=\tilde{N}_j, D(t_j)=\tilde{D}_j\ \forall j\in\{1,\dots,N\}\}$. Since we assume that the deterministic growth process for the primary tumour is uniquely defined by the corresponding growth rate parameters in $\boldsymbol{\theta}$ the observation likelihood is independent from the realized path and we can then rewrite

$$\begin{aligned}&\int_{\Omega_{\boldsymbol{\theta}}}\prod_{j=1}^{N}p_{\boldsymbol{\theta}}(\mathbf{y}_j|\mathbf{X}(t_j))p_{\boldsymbol{\theta}}(\mathbf{X}(t_j)|\mathbf{X}_{t_{j-1}})d\mathbf{X}\\ &\quad=\prod_{j=1}^{N}p_{\boldsymbol{\theta}}(\mathbf{y}_j|S(t_j))\int_{\tilde{\Omega}_{\boldsymbol{\theta}}}\prod_{j=1}^{N}p_{\boldsymbol{\theta}}(\mathbf{X}(t_j)|\mathbf{X}_{t_{j-1}})d\mathbf{X}\end{aligned}\tag{12}$$

Given the initial condition N_0, D_0 the second part of the likelihood (12), the process likelihood, is given by

$$\begin{aligned}\int_{\tilde{\Omega}_{\boldsymbol{\theta}}}\prod_{j=1}^{N}p_{\boldsymbol{\theta}}(\mathbf{X}(t_j)|\mathbf{X}_{t_{j-1}})d\mathbf{X} &= \prod_{j=1}^{N}\mathbb{P}(N(t_j)=\tilde{N}_j, D(t_j)=\tilde{D}_j|N_{t_{j-1}}, D_{t_{j-1}})\\ &= \prod_{j=1}^{N}\mathbb{P}(D(t_j)=\tilde{D}_j|N(t_j), D_{t_{j-1}}, N_{t_{j-1}})\,\mathbb{P}(N(t_j)=\tilde{N}_j|N_{t_{j-1}}, D_{t_{j-1}})\\ &= \prod_{j=1}^{N}\mathbb{P}(D(t_j)=\tilde{D}_j|N(t_j), D_{t_{j-1}}, N_{t_{j-1}})\,\mathbb{P}(N(t_j)=\tilde{N}_j|N_{t_{j-1}}),\end{aligned}$$

which decomposes the likelihood of the process dynamics into a likelihood contribution of the metastasis counting process and one of the death process. Both are investigated in more detail in the next steps. Note that on the right hand side and in the rest of the section we omit the dependence on $\boldsymbol{\theta}$ for notational brevity.

Likelihood of the observed number of metastasis

Since the metastasis number is not bounded from above and we are assuming an underlying time-inhomogeneous Poisson counting process, the likelihood contribution

of the metastasis process is given by

$$\begin{aligned}\mathbb{P}(N(t_j) = n_j | N(t_{j-1}) = n_{j-1}) \\ = \frac{\Lambda_N([t_{j-1}, t_j), S)^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} \exp\left(-\Lambda_N([t_{j-1}, t_j), S)\right),\end{aligned} \tag{13}$$

where S is the deterministic tumour growth process, i.e. the solution of (1) uniquely defined by the model parameters and the initial condition corresponding to the tumour growth.

Likelihood of the observed timepoint of death

For the likelihood contribution of the death process, we need to marginalise out the unseen timepoints at which new metastases occur between two observation timepoints t_{j-1}, t_j, leading to the following results.

Theorem 1 (Survival likelihood) *Let $S(t)$ be the solution to the ODE (1) satisfying Assumption 1; let $\lambda_N(t, S(t))$ be the intensity function of a Poisson point process $N(t)$ admitting Assumption 2; let $\lambda_D(t, S(t), N(t))$ be the intensity function of $D(t)$ given by (7). Then the probability of survival given the occurrence of m new jumps of $N(t)$ in a time interval $[t_{j-1}, t_j)$ is given by*

$$\begin{aligned}&\mathbb{P}(D(t_j) = 0 | D(t_{j-1}) = 0, N(t_{j-1}), N(t_j)) \\ &= \int_{t_{j-1}}^{t_j} \int_{u_1}^{t_j} \cdots \int_{u_m}^{t_j} \frac{m! \prod_{i=1}^m \lambda_N(u_i, S(u_i))}{\Lambda_N([t_{j-1}, t_j), S)^m} \\ &\quad \prod_{i=0}^{m} \exp\left(-\int_{u_i}^{u_{i+1}} \lambda_D(s, S(s), N(s)) ds\right) du_m \dots du_1,\end{aligned} \tag{14}$$

where $u_0 := t_{j-1}$ and $u_{m+1} := t_j$.

Using this, one can directly deduce the probability of death occurring at time t_j given that m new metastasis occurred after the last observation at t_{j-1}.

Corollary 2 *(*Death likelihood) *Under the same assumptions as before, we get*

$$\begin{aligned}&\mathbb{P}(D(t_j) = 1 | D(t_{j-1}) = 0, N(t_{j-1}), N(t_j)) \\ &= \int_{t_{j-1}}^{t_j} \int_{u_1}^{t_j} \dots \int_{u_m}^{t_j} \frac{m! \prod_{i=1}^m \lambda_N(u_i, S(u_i))}{\Lambda_N([t_{j-1}, t_j), S)^m} \\ &\lambda_D(t_j, S(t_j), N(t_j)) \prod_{i=0}^{m} \exp\left(-\int_{u_i}^{u_{i+1}} \lambda_D(s, S(s), N(s)) ds\right) du_m \dots du_1.\end{aligned} \tag{15}$$

The proofs of these two results can be found in the Supplementary Information D.

Analytical likelihood formulation for models with an exponential tumour growth law. For models based on an exponential tumour growth law as described in Sect. 2.1.1 combined with any of the two metastasis processes and the death process presented in Sect. 2, we were able to solve the nested integrals involved in Eqs. (14) and (15) analytically. This yields a completely analytically available likelihood function for these models. The precise computation and simplification of the analytical solutions were done using symbolic computation in the Wolfram Mathematica computer algebra system (Wolfram Research, Inc. 2024).

For the model choices for which an analytical solution was not achievable, we leveraged efficient numerical integration schemes to solve the nested integrals and obtain a numerical approximation to the likelihood function that could then be further used for the optimisation (Leader 2022).

Likelihood calculation in the case of missing data Models designed for the use with clinical data face several challenges, notably the use of irregularly sampled, noisy and partially missing measurements. The first is addressed by the use of different observation noise models as pointed out in Sect. 2.2. Corresponding to the data at hand such noise models can be applied to each of the subprocesses of the combined model. Irregularly sampled data are naturally supported by the model as it does not assume any structure on the distribution of measurement times and the likelihood for the proposed modelling framework can be factorized over timepoints, see Eq. (11), this is inherently addressed and one of the advantages of a Markovian model. Partially missing timepoints, where only measurements of some of the subprocesses are provided, can be naturally addressed as well, by marginalising over the missing components of the model. Let the measurement of the i-th component of the process X be missing at timepoint t_j, then the likelihood reads as

$$\prod_{k=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{X}(t_k)|\mathbf{X}_{t_{k-1}}) = \prod_{k\notin\{j,j+1\}} p_{\boldsymbol{\theta}}(\mathbf{X}(t_k)|\mathbf{X}_{t_{k-1}}) \int_{\Omega_\theta^i} p_{\boldsymbol{\theta}}(\hat{\mathbf{X}}^i(t_j)|\mathbf{X}(t_{j-1}))\, p_{\boldsymbol{\theta}}(\mathbf{X}(t_{j+1})|\hat{\mathbf{X}}^i(t_j))\, d\hat{x}, \quad (16)$$

where $\hat{\mathbf{X}}^i(t_j)$ denotes the process with the i-th component being $X^i(t_j) = \hat{x}$.

Given that, if a measurement timepoint exists, one knows at least the survival status of the patient we consider the case of missing tumour size measurement and missing metastasis number measurement. In the first case, marginalising over all possible tumour sizes will just yield an likelihood value of the observation model equal to 1. Hence, we can just omit the likelihood of the observation model at the timepoint t_j.

In the case of a missing metastasis number measurement, the integral in (16) will become a sum.

If the patient dies at timepoint t_j, i.e. $\tilde{D}_j = 1$, we can write

$$\mathbb{P}(D(t_j) = \tilde{D}_j | N(t_j), D_{t_{j-1}}, N_{t_{j-1}}) \mathbb{P}(N(t_j) = \text{missing} | N_{t_{j-1}})$$

$$\approx \sum_{m=0}^{M} \mathbb{P}(D(t_j) = \tilde{D}_j | N(t_j), D_{t_{j-1}}, N_{t_{j-1}}) \mathbb{P}(N(t_j) = m | N_{t_{j-1}})$$

where the cut-off M is chosen to be the highest number of metastasis we observed in the data.

If the patient does not die at t_j the marginalised likelihood will be

$$\begin{aligned}
&\mathbb{P}(D(t_j) = \tilde{D}_j | N(t_j), D_{t_{j-1}}, N_{t_{j-1}}) \mathbb{P}(N(t_j) = \text{missing} | N_{t_{j-1}}) \\
&\mathbb{P}(D(t_{j+1}) = \tilde{D}_{j+1} | N(t_{j+1}), D_{t_j}, N_{t_j}) \mathbb{P}(N(t_{j+1}) = \tilde{N_{j+1}} | N_{t_j}) \\
&\quad = \sum_{m=0}^{\infty} \mathbb{P}(D(t_j) = \tilde{D}_j | N(t_j), D_{t_{j-1}}, N_{t_{j-1}}) \mathbb{P}(N(t_j) = m | N_{t_{j-1}}) \\
&\mathbb{P}(D(t_{j+1}) = \tilde{D}_{j+1} | N(t_{j+1}), D_{t_j}, N_{t_j}) \mathbb{P}(N(t_{j+1}) | N_{t_j}) \qquad (17)
\end{aligned}$$

where all summands with $m \notin [\tilde{N}_{j-1}, \tilde{N}_j]$ are zero.

For the case where the patient does not die in the interval $[t_{j-1}, t_{j+1}]$ Eq. (17) simplifies even more. Leveraging the Chapman-Kolmogorov equation (Pavliotis 2014) one can just skip the process likelihood at timepoint t_j and at the next timepoint compute $p_{\boldsymbol{\theta}}(\mathbf{X}(t_{j+1} | \mathbf{X}(t_{j-1})$.

3.2 Parameter optimisation

The maximum likelihood estimate $\boldsymbol{\theta}^{\text{MLE}}$ of the model parameters is obtained by maximising the likelihood of observing the data given the model, computed in the sections above. In this work, we determine the maximum likelihood estimate by minimising the negative log-likelihood yielding

$$\boldsymbol{\theta}^{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} - \ln\left(\mathcal{L}(\boldsymbol{\theta} | \mathcal{D})\right),$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters characterizing the combined stochastic process $\mathbf{X}$. The minimization is done with computational optimisation techniques that need to be chosen according to the likelihood formulation at hand. For the case of analytical likelihood functions, we additionally compute gradients to leverage a gradient-based minimisation algorithm such as gradient descent or quasi-Newton methods, whereas with the numerical approximation to the likelihood we fall back to non-gradient based methods such as simulated annealing (Press 2007).

3.3 Uncertainty analysis

For models which provide an interpretation of the model parameters, an assessment of how certain we are in the parameter estimates is meaningful. We assessed this parameter uncertainty by approximating the posterior distributions of the parameters

$p(\boldsymbol{\theta}|\mathcal{D})$. This was done using a Markov chain Monte Carlo (MCMC) sampling algorithm (Andrieu et al. 2003), to approximate the posterior distribution by a large number of samples and compute corresponding credibility intervals for $\boldsymbol{\theta}^{\mathrm{MLE}}$. More precisely, we used an adaptive parallel tempering algorithm (Vousden et al. 2015) to run several independent Markov chains and discarded the first half of the chain as a burn-in.

3.4 Implementation of the inference pipeline

As the code for the model simulation, we also implemented the likelihood-based inference procedure purely in the Julia programming language. Analytical solutions of the likelihoods were written down precisely and gradients were, if possible, computed by automatic differentiation (Revels et al. 2016). For the numerical approximations of the likelihood we applied the numerical integration schemes from the `Cubature.jl` package (Johnson 2020). In particular, we leveraged an p-adaptive numerical integration scheme based on Clenshaw-Curtis quadrature rules Clenshaw and Curtis (1960). This can be seen as an expansion of the integrand in terms of Chebyshev polynomials and is well suited for accurately calculating low-dimensional integrals (Johnson 2010). We chose a relative error tolerance of $1e^{-8}$. While in principle differentiable with respect to the model parameters, the implementation using `Cubature.jl` for a robust and fast evaluation of the nested integrals does not allow for differentation.

Following comprehensive testing, we decided for running multiple starts of a local optimisation algorithm. For gradient-based optimisation in the cases where the gradient was available, we employed the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm Liu and Nocedal (1989). For gradient-free optimisation, we used the simulated annealing with box constraints (SAMIN) algorithm (Kirkpatrick et al. 1983; Goffe 1996). Both are implemented in the `Optim.jl` package (Mogensen and Riseth 2018). For improved numerical stability, model parameters were transformed to log-scale during the optimisation.

For Markov chain Monte Carlo sampling, we used an adaptive parallel tempering algorithm (Vousden et al. 2015) as implemented in the Python package `pyPESTO` (Schälte et al. 2023) that was interfaced from Julia.

All computations were conducted on 10 cores of an AMD EPYC 7F72 3.2 GHz processor with 20GB of RAM. In order to ensure reusability and reproducibility, we made the code for model simulations and experiments together with the artificial data and results generated for this paper available at Zenodo (https://doi.org/10.5281/zenodo.13839104).

4 Evaluation of parameter inference

In this manuscript, we introduced a model class for the joint analysis of tumour growth, metastatic development, and patient survival. In the following, we assess the computational complexity of likelihood evaluation and parameter estimation. Furthermore, we evaluate the agreement of numerically and analytically computed likelihood values, the accuracy of parameter inference for different model structures, and the distin-

Table 1 Overview of the models used for the simulation study in Section 4

Model name	Tumour growth process $S(t)$	Metastasis process intensity function
(M1) Exponential proportional model	$S(t) = S_0 \exp(\beta t)$	$\lambda_N^{\text{prop}}(t) = m_{\text{basal}} + m_{\text{size}}\sqrt{S(t)}$
(M2) Cell division model	$S(t) = S_0 \exp(\beta t)$	$\lambda_N^{\text{div}}(t) = m_{\text{division}}\beta\left(\frac{\beta t}{\ln(2)}\right)$
(M3) Gompertz model	$S(t) = K \exp\left(\ln\left(\frac{S_0}{K}\right)\exp(-\beta t)\right)$	$\lambda_N^{\text{prop}}(t) = m_{\text{basal}} + m_{\text{size}}\sqrt{S(t)}$
(M4) Gyllenberg–Webb model	$S(t) = Q(t) + P(t) + De(t)$	$\lambda_N^{\text{prop}}(t) = m_{\text{basal}} + m_{\text{size}}\sqrt{S(t)}$
(M5) Treatment effect model	$S(t) = S_0 \exp(\beta \mathbf{Z}(t)t)$	$\lambda_N^{\text{prop}}(t) = m_{\text{basal}} + m_{\text{size}}\sqrt{S(t)}$

guishability of different parameterisations and model structures. Overall, we consider 5 different example models consisting of different combinations of growth and metastasis processes as summarized in Table 1. The death process is the same for each of the models.

4.1 Study design

For each of the models, we created a dataset of 500 patients under the following experimental conditions. Using the corresponding model description, we simulated the whole trajectory until the death of the patient or a maximum time of 30 months and read out observations at timepoints $t_0, \ldots, t_n$, corresponding to monthly visits of the patient.

For the simulation of the tumour growth given by (1) we assumed that tumours originate from a single spherical cell of volume S_{cell} corresponding to a diameter of $d_{\text{cell}} = 0.01$ mm, but are detectable only once they have reached a volume $S_{\text{detection}}$ corresponding to a diameter of $d_{\text{detection}} = 0.5$ mm (Gasparini and Humphreys 2022; Hao et al. 2018). Hence, the first observation timepoint t_0 of a patient corresponds to the time between offset of the tumour and diagnosis. However, under the assumptions that the initial volume and detection threshold are known and given the model parameters, we could calculate the time since the offset of the tumour t_0 for each patient, based on the first observation. Therefore, without loss of generality we assumed that tumours are detected as soon as they reach the detection threshold, i.e. $S_0 = S_{\text{detection}} = 0.065\,\text{mm}^3$ for all patients and omitted the constant linear time shift and set $t_0 = 0$ for all patients. Additionally, the Markov property of the metastasis process ensures that the gain of new metastasis is independent of how many metastasis the patient has developed so far. Therefore, without loss of generality we initialized the patient simulations with zero metastasis at the first observation timepoint, i.e. $N_0 = 0$.

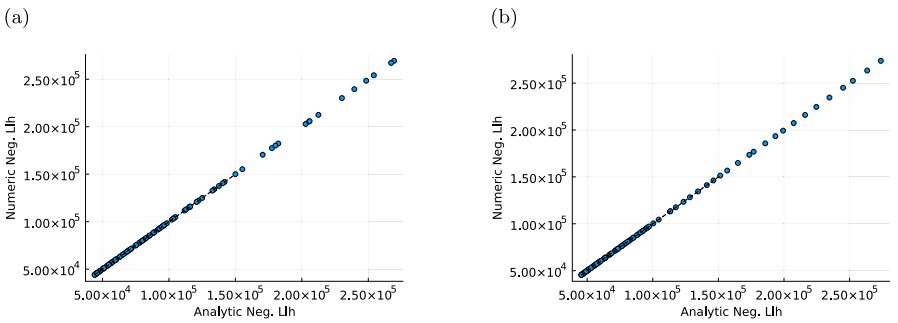


Fig. 6 Comparison of likelihood formulations. We randomly sampled 100 parameter vectors and evaluated the analytical negative log-likelihood as well as the numerical negative log-likelihood on this vectors. **a** Visualizes the resulting values for the exponential proportional model and **b** for the cell division model. The black-dashed lines correspond to the 45° lines

4.2 Numerical and analytical likelihood values agree for exponential growth based models

In Sect. 3.1, we formulated the likelihood function and described the availability of analytical solutions for certain model formulations. Here, we examine the solution of the likelihood function for model (M1) and (M2), which are based on the exponential tumour growth law. The combination of the exponential tumour growth law with the volume-based metastasis intensity enables the computation of the accumulated metastasis intensity rate as

$$\Lambda_N^{\text{prop}}([t_{j-1}, t_j)) = (t_j - t_{j-1}) m_{\text{basal}} + \frac{2 m_{\text{size}} \sqrt{S_0}}{\beta} \left(\sqrt{\exp(\beta t_j)} - \sqrt{\exp(\beta t_{j-1})} \right).$$

For the second model the use of the cell division based intensity rate for the metastatic spread (6), yields

$$\Lambda_N^{\text{div}}([t_{j-1}, t_j)) = \frac{m_{\text{division}}}{(k+1)\ln(2)^k} \beta^{k+1} \left(t_j^{k+1} - t_{j-1}^{k+1} \right).$$

These accumulated metastasis intensities directly provide analytical expressions for the metastasis likelihood (13). Moreover, both model formulations enable the computation of analytical solutions for the nested integrals in the likelihood contribution of the death process Eqs. (14) and (15). The precise expressions for the number of new metastasis $m \in \{1, 2, 3, 4, 5\}$ are given in the Supplementary Information E, where the upper bound on m was chosen due to the maximal number of new metastasis observed in one inter-observation time interval in the simulated datasets. Artificial data was simulated using parameters provided in Supplementary Table S2

For both models evaluating the resulting analytical negative log-likelihood function shows a complete agreement with the corresponding approximated negative log-likelihood function obtained by numerical integration as seen in Fig. 6.

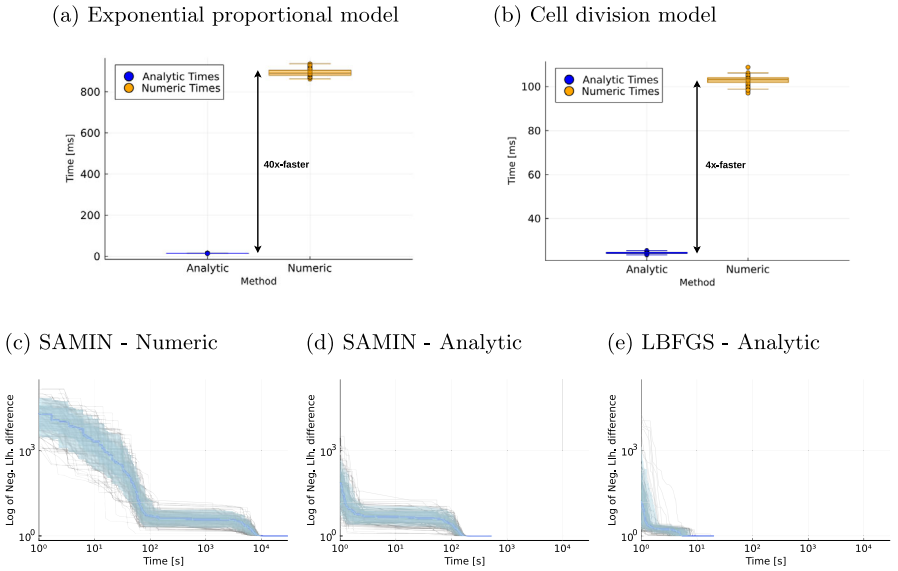


Fig. 7 Evaluation of computational efficiency. **a**, **b** show the mean computation times of evaluating the negative log-likelihood function on 100 randomly sampled parameter vectors for the two exponential growth based models (M1) and (M2). **c**, **d**, and **e** depict the trace of the current best negative log-likelihood value over time for 100 single optimisation runs of the exponential proportional model using different likelihood formulations and optimization algorithms

4.3 Efficiency gain by using analytical likelihoods

Next, we investigated the computation time for evaluating the analytical and numerical negative log-likelihood functions for the two exponential tumour growth based models (M1) and (M2). For this, we repeatedly evaluated each function on 100 randomly sampled parameter vectors and stored the mean computation time for each parameter vector.

For both models, evaluation of the analytical likelihood was computationally more efficient than for their numerically approximated counterpart (Fig. 7). In the exponential proportional model, we even observe a more than fifty-fold speed-up. In the cell division model, the evaluation is around four times faster for the analytical likelihood. The precise mean computation times for the evaluation can be found in the Supplementary Information Table S3.

To assess the daily applicability of the combined model for analysing patient trajectories from clinical data, we are interested in the computation time of optimizing the negative log-likelihood of the models given the data. For each model we sampled 100 parameter vectors as start points for optimisation and in each optimisation run traced the current best negative log-likelihood value over time. We observed a four-time faster optimisation in the cell division model by using the analytical likelihood function instead of numerical approximations in the simulated annealing optimisation algorithm. Leveraging automatic differentiation, we computed gradients for the analytical likelihood functions, their second big advantage, and were able to use a gradient-based

optimisation algorithm. This sped-up optimisation by another 100-times (Supplementary Figure S3). For the exponential proportional model computation time decreased fifty-fold by using analytical instead of numeric likelihoods. Using the gradient based optimiser for the local optimisation task, the efficiency gain was about another forty-fold, as depicted in Fig. 7. In total, this resulted in a 2,000 times faster optimisation and shows the benefit of analytical likelihoods and their gradients whenever possible. The precise mean computation times for the optimisation runs can be found in the Supplementary Information Table S4.

Moreover, we observed that the likelihood formulation did not influence the capability of the optimiser to converge to the optimum and convergence of the runs shows great reproducibility of the optimisation (Supplementary Information Figure S7).

4.4 Analytical likelihoods facilitate accurate parameter inference

The proposed modelling framework aims to identify rates that determine the disease dynamics for individual patients. Therefore, we need to assess the uncertainty of the parameter estimates. The uncertainties were computed using MCMC methods. We found that the credibility intervals cover the parameters used to generate the simulated data. For the exponential proportional model (M1) the true value lies even in the smallest credibility interval of 80% for three parameters, showing a high accuracy of the estimation procedure (Fig. 8). The influence of the metastasis on the death process, given by $d_{\text{metastasis}}$, has the highest variance, as the result of the dependence of this parameter on two stochastic processes. The highest relative deviation between true and estimated parameter is in the base rate of the metastatic spread given by m_{basal}.

In the cell division model (M2), only one parameter m_{division} determines the trajectory of the metastasis process, reducing the uncertainty in the metastasis rate. Hence, for all four model parameters, the true values are covered not just by the 95%, but also by the 90% and 80% credibility intervals (Fig. 9). This reveals the capabilities of the model to retrieve the underlying parameter values with high certainty. As before, the highest variance is in $d_{\text{metastasis}}$ due to its stochastic nature. The high uncertainty in that parameter is also reflected in the estimated density of its marginal likelihood. In contrast to the other parameters, it admits a right-skewed distribution with a higher variance (Supplementary Information Figures S8,S9).

In clinical application, the data at hand does not consist of equidistant and structured observations of the patients. To assess the performance of the proposed modelling and inference framework, we ran the optimisation for the exponential proportional model on a dataset including completely and partially missing timepoints. To be precise, we randomly removed about 30% of all measurement timepoints from the complete dense dataset with monthly observations. Additionally, we assigned about 20% of the tumour measurements and about 20% of the metastasis measurement to be missing. Using this sparse and dataset with partially missing timepoints, we still obtained accurate estimates $\boldsymbol{\theta}^{MLE}$ and credibility intervals covered the parameters used for data generation (Supplementary Information Figure S4). As expected the credibility intervals showed more uncertainty in the precise parameter values due to a less informative dataset.

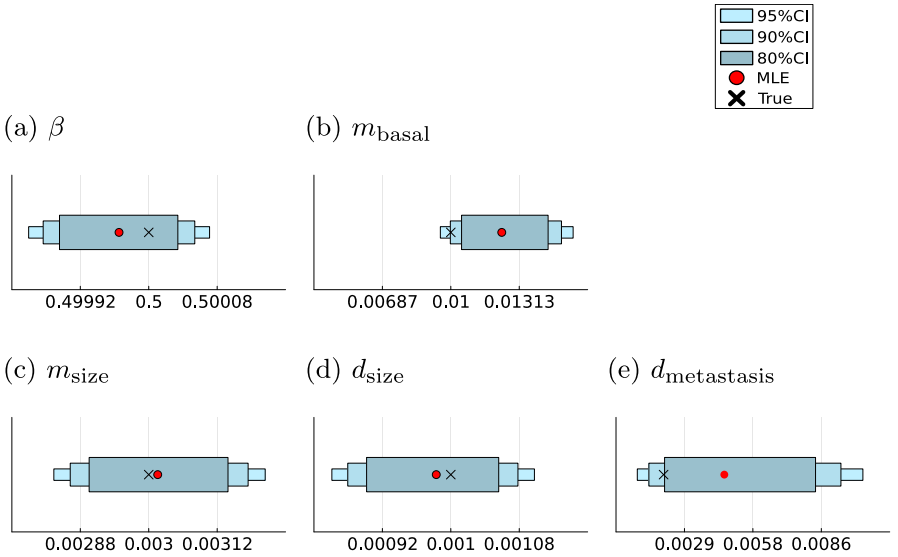


Fig. 8 Parameter inference results for model (M1) Maximum likelihood estimates (MLE) and sampling-based credibility intervals for the model parameters of the exponential proportional model

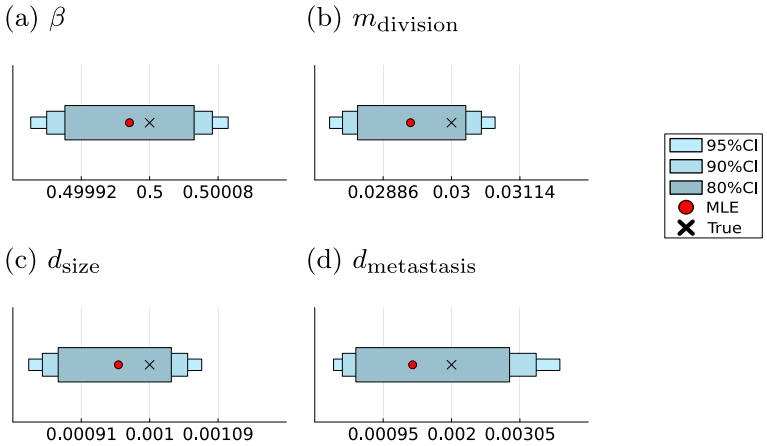


Fig. 9 Parameter inference results for model (M2) Maximum likelihood estimates (MLE) and sampling-based credibility intervals for the model parameters of the cell division model

4.5 Numerical likelihood approximations enable flexible model formulations

One further advantage of the proposed modelling framework is its flexibility. Yet, in the models considered so far, we only included models based on exponential growth, which allowed us to derive analytical likelihood formulations. However, the use of numerical integration to approximate the likelihood function allows for more flexible choices of the different processes. As examples for more complex tumour growth we consider the Gompertz growth model given in (3) and the Gyllenberg–Webb model

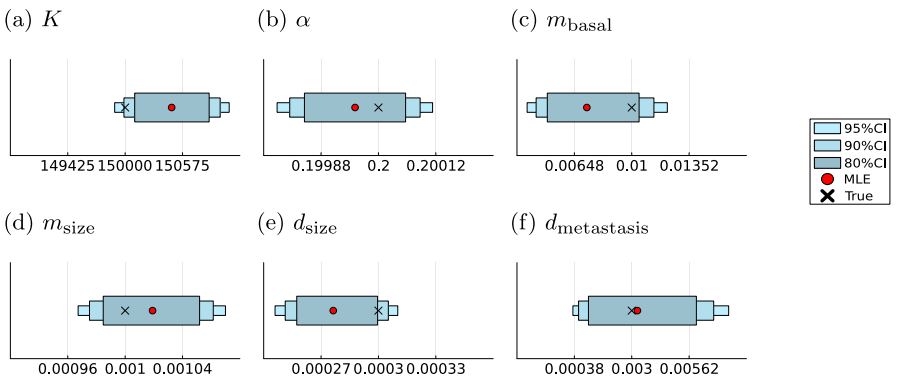


Fig. 10 Parameter inference results for model (M3) Maximum likelihood estimates (MLE) and sampling-based credibility intervals for the model parameters of the Gompertz model

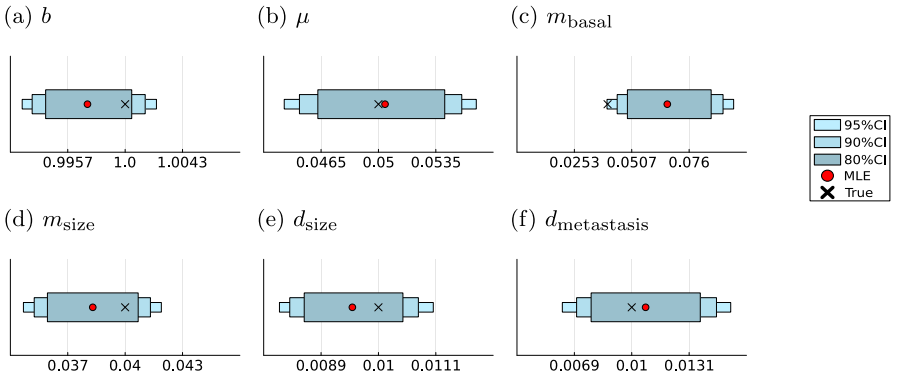


Fig. 11 Parameter inference for model (M4). Maximum likelihood estimates (MLE) and sampling-based credibility intervals for the model parameters of the Gyllenberg–Webb model

based on Eq. (4). Both are combined with the volume based intensity rate function for the metastasis process as given in equation (5). For the Gompertz model we estimated the growth parameters $\boldsymbol{\theta}_{\text{Gompertz}}[1,2] = (K, \alpha)^T$ and for the Gyllenberg–Webb model the proliferation and death rate $\boldsymbol{\theta}_{\text{GW}}[1,2] = (b, \mu)^T$. All other parameters for the Gyllenberg–Webb model were fixed to the example given in (Alzahrani et al. 2014). We generated artificial data from the resulting models (M3) and (M4) using the model parameters provided in Table S7.

Subsequently, we performed 100 starts of minimizing the negative log-likelihood initialized at randomly sampled startpoints. For both models all starts converged to the same minimum (Supplementary Figure S11), yielding maximum likelihood estimators $\boldsymbol{\theta}^{\text{MLE}}_{\text{Gompertz}}$ and $\boldsymbol{\theta}^{\text{MLE}}_{\text{GW}}$ for the model parameters. Furthermore, the credibility intervals obtained by MCMC methods cover the parameter values used to generate the data as shown in Figs. 10 and 11. In the Gompertz model all parameters even lie in the the 90% credibility interval showcasing high certainty in the estimates. And again the highest variance is visible in the doubly stochastic parameter $d_{\text{metastasis}}$. In the Gyllenberg–

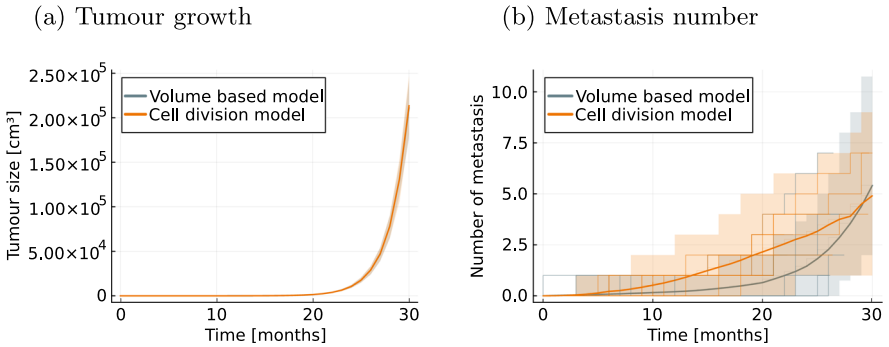


Fig. 12 Datasets for model selection. Visual comparison of the two datasets used for the model selection task

Webb model the base metastasis rate shows the highest uncertainty. However, all parameters are contained in the 95% credibility intervals.

4.6 The analysis framework allows to distinguish between different metastasis processes

In biomedical studies the structure of the underlying processes is often partially unknown. In this cases, competing hypotheses can be compared using model selection. This can be achieved by comparing the Akaike information criterion (AIC) (Akaike 1974) of competing model formulations that try to explain the data. The AIC estimates the quality of each model by balancing between the goodness of fit and the simplicity of the model

$$\mathrm{AIC} = 2 \cdot p - 2 \cdot \hat{l}, \tag{18}$$

where p denotes the number of model parameter and $\hat{l} = \ln(L(\boldsymbol{\theta}^{\mathrm{MLE}}))$ the maximum value of the log-likelihood function.

Here, we assess the ability of the proposed framework to determine the correct model structure, by creating artificial data for the two exponential tumour growth based models (M1) and (M2) (Fig. 12) using parameters provided in Table S2. The metastasis processes differ in their dynamics and show a exponential like increase in the number of metastases for the exponential proportional model with volume based metastasis intensity function and a linear increase for the model with the cell division based metastasis intensity function (Fig. 12).

Each of the resulting datasets was subsequently fitted using both models the AIC values were computed. We observed that the AIC is minimal if the model coincides with the process underlying the data and we could then identify the correct model formulation. Furthermore, the difference in AIC value was significant for the wrong model formulation yielding high confidence in the selected model (Table 2). Similarly, the concordance between model and data is displayed when plotting the fit of

Table 2 Model selection results

Model	Data	
	Volume based	Cell division
Volume based	0.0	1141.927
Cell division	354.555	0.0

Difference of AIC values for the model given the dataset and the best AIC value obtained for that dataset

observations against model simulations model obtained with the estimated parameter (Supplementary Information Figure S10).

4.7 The model supports identification of treatment and covariate effects

In addition to modelling individual patient trajectories, another goal in clinical research is the identification of treatment effects and covariates responsible for changes in the dynamics of the disease. Here, we extended the exponential proportional model by introducing a therapy effect that reduces the growth of the tumour as explained in Sect. 2.1.5. Additionally, we introduce the obesity status of the patient as a covariate that effects the therapy, where obesity is defined as a Body Mass Index greater or equal 30 (World Health Organization 2024). Obesity has been shown to be closely related to the development, characteristics and treatment efficacy of several cancer entities (Ross et al. 2019; Wolin et al. 2010). The realisation of treatment and obesity for a patient l are stored in the individual covariate vector

$$\mathbf{Z}^l(t) = \begin{pmatrix} \mathbb{1}_{\text{Treatment at time } t} \\ \mathbb{1}_{\text{ Obese at time } t} \end{pmatrix}.$$

where we assume that the obesity status does not change over time and once a treatment is in place, it will be given until the end of the observations for that patient.

This results in the following functional form of the tumour growth rate

$$\beta(t, \mathbf{Z}^l) = \beta_0 + \mu^T \mathbf{Z}^l(t),$$

where $\mu^T = (-\rho, \delta)$ is the vector of effects of the covariates on the basal growth rate β_0 and encodes the presence of a treatment and the obesity status of the patient l at time t. We generated artificial data using parameters given in Table 3 in which the treatment start time and the obesity status were randomly drawn. Both covariates are assumed to be observed directly. We were able to reuse the computation of the analytical likelihood for the standard exponential proportional model to retrieve analytical likelihood for this model as well. Subsequently, we fitted the model running 100 starts of the LBFGS optimisation algorithm to obtain the MLE for the resulting model parameter vector $\boldsymbol{\theta} = (\beta_0, \rho, \delta, m_{\text{basal}}, m_{\text{size}}, d_{\text{size}}, d_{\text{metastasis}})^T$. All runs converged to the same optimal point (Supplementary Information Figure S13). The resulting estimates accurately recover the treatment effects and credibility intervals obtained by MCMC methods cover the parameters underlying the data (Fig. 13). Opposed to

Table 3 Treatment effect model parameterization

Parameter	True value	Parameter range
β	-0693	$[-0.75, -0.65]$
ρ	-0.799	$[-0.85, -0.75]$
δ	-1.609	$[-2.3, -1.2]$
m_{basal}	-4.605	$[-7, -2]$
m_{size}	-6.908	$[-7, -2]$
d_{size}	-8.1117	$[-9, -4]$
$d_{\text{metastasis}}$	-5.809	$[-9, -4]$

Model parameters with their corresponding ranges on log-scale for the treatment effect model

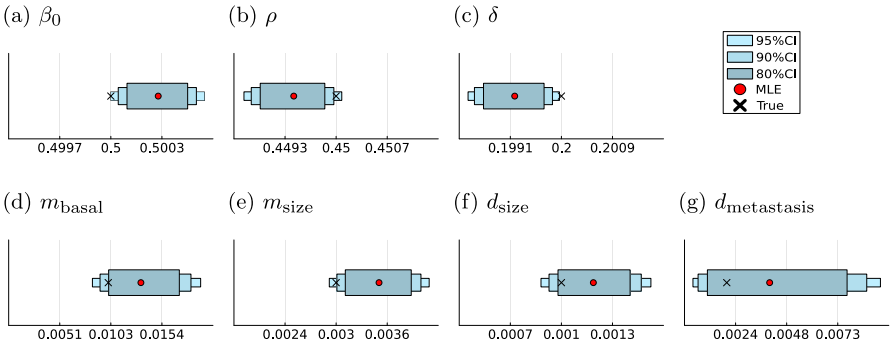


Fig. 13 Parameter inference for model (M5) Maximum likelihood estimates (MLE) and sampling-based credibility intervals for the model parameters of the treatment effect model

the exponential proportional model the tumour growth parameters are not covered by the 80% credibility interval anymore, but by the 95% interval. However, the model still shows high certainty in this parameters shown by the small variance of these parameters and hence small credibility intervals.

5 Discussion

The proposed class of combined stochastic models provide a useful alternative to multi-state or continuous growth models for studying cancer progression. While there exist many models that model tumour growth and metastatic seeding in special cancer entities, e.g. for breast cancer (Gasparini and Humphreys 2022), none of them directly couples this with the survival of the patients, or is flexible enough to be applied to different cancer entities. The modelling framework proposed in this paper explicitly combines the health status of the patient and the disease progression. Thus, it provides insights into the tumour growth and spread processes over time. It is the first of its kind and a first step towards holistic models for the analysis of patient trajectories and the development of practicable models for cancer progression.

The models we use in this work are examples of concrete model parameterisations. Yet, they still inherit some simplifying assumptions limiting their applicability in practice. In addition to the assumptions made in Sect. 2, we only applied them to simulated data, which allowed some simplifying assumptions on the initial point S_0 and t_0. When applied to real data, every patient would exhibit a different value for S_0. However, given the model formulation, we can then compute the time until detection t_0 and shift the time scale for each patient individually.

Moreover, the use of artificial data yields a more structured environment. We have shown that time intervals between observations need not to be equidistant and have unregular data sets. Therefore, we adapted the formulation of the likelihood to time points being unstructured and without complete observations, i.e. only some of the subprocesses are observed. Although, this does not fully resemble the complexity of data collected in clinical routine, it proves that the concept of the modelling framework can easily be applied to various realistic datasets.

In addition, although patients could differ in the realisation of their covariates in Sect. 2.1.5, leading to individual trajectories, we only considered the case where all model parameters are the same for all patients. This issue can be addressed by introducing random differences in the parameters across patients by for example the use of a mixed effect like description for the tumour growth rate.

Moreover, to take advantage of all the information contained in the patient trajectories, the processes involved in the model should include more covariates, such as biomarkers or general patient characteristics. Time-varying covariates can also account for resistance to cancer therapy over time. Additionally to introducing covariates on the cancer growth rates, one can also add covariates to the rates of the other subprocesses. Such adjustments are depending on the data and question at hand, but the proposed modelling framework can serve as a good exploratory tool. Therefore, future work on possible rate parametrizations and extensions of the likelihood formulations is encouraged.

Although, we consider here only simple models for the subprocesses, the proposed framework provides a high degree of flexibility. While our results demonstrate that already the considered process descriptions captured real-world data, model refinements which capture additional biological and clinical details should be explored, e.g., the study design and the training process to avoid model bias. The precise model formulation should depend on the prior knowledge about the cancer entity, the available data and the research question.

One simplifying assumption we made in this manuscript, is to only consider exponentially distributed survival times. As shown in Sect. 2.4 this does not limit the overall survival times to be exponentially distributed. However, further enhancements by considering more realistic survival time distributions, such as a Weibull or a Gamma distribution, can be of interest in ongoing research.

Another pathway of future extensions is to consider even more complex tumour growth and metastasisation models that can for example capture spatial dynamics and micrometastases. Using PDE based models for the tumour growth, such as the Greenspan model (Greenspan 1972; Bull and Byrne 2022), can be incorporated similarly to the physiologically structured Gyllenberg–Webb model (Gyllenberg and Webb 1990; Kuang et al. 2018) by adapting the likelihood formulation. Altering the metas-

tasisation process to a finer scale or even using a progression model for the tumour growth and metastasisation on a cellular level (Rupp et al. 2024) can lead to more insights.

Our proposed modelling framework is mainly concerned with modelling the patient trajectories on a higher level using clinical routine data. We did combine processes acting on different scales by using a physiologically structured tumour growth model with an metastasis and survival process dependent only on the total tumour size. However, this does not consider the whole multi-scale nature of cancer yet. Models of this important aspect of cancer and for example account for micrometastases as in L Rocha et al. (2024) would need more fine grained data. Additionally, this would need further investigation on the formulation of the corresponding likelihood function and is out of the scope of this work. But we highly encourage further research in providing a combined multi-scale model.

While the flexibility of the modelling framework allows for building and simulation of such extended and more complex model parameterisation, another bottleneck lies in the efficiency of the parameter estimation procedure. If the formulated model exhibits an analytical likelihood formulation, maximum likelihood estimation can be done efficiently. Otherwise, approximating the likelihood or using likelihood-free and simulation based Bayesian inference schemes can lead to inefficient estimation tasks. Improvement of those for this class of models is out of the scope of this paper and a direction for following research.

In this paper, we leveraged automatic differentiation for the computation of gradients and gradient-based optimisation for the models based on an exponential tumour growth law. However, further investigation into the computation or numerical approximation of the likelihood can lead to differentiable objective functions for more models. This could then allow the use of gradient-based optimisation and a higher computational efficiency for those models as well. Depending on the model formulation and its observables, it is also possible to estimate parameters not jointly, but in a hierarchical manner. For example, in the cell division model with all three processes being observed, one could use the analytical and differentiable expressions for the tumour and metastasis rate parameters to estimate them with an efficient gradient-based method. The resulting estimates may then be fixed in the death process likelihoods, which are then optimized with the non-gradient based method. This can further improve the optimisation procedure. Hence, the inference procedure is a subject of ongoing research as well and could profit from further advances in computational frameworks and software packages.

One further direction for developing a more realistic description of cancer dynamics can be to evolve to a completely stochastic model by exchanging the deterministic growth laws for the tumour growth by their corresponding stochastic counterpart or a general SDE representation. More recent models suggest this exchange by stochastic differential equations for tumour growth (Ayuni Mazlan et al. 2017; Mansour and Abobakr 2022; Katsaounis et al. 2023). This then poses new challenges on computing or approximating the likelihood function and hence, the ability for efficient parameter inference. Therefore, further research on numerical approximations for such extended models or the use and efficiency of Bayesian estimation frameworks in this context can provide interesting new ways to extend this modelling framework further.

The combined stochastic model can address all six hallmarks of mathematical oncology (Bull and Byrne 2022). Moreover, the flexibility of the framework allows it to emphasize different aspects of the progression of cancer. As discussed before a tumour growth and metastasisation model could be used on a much more detailed scale, while keeping the survival process more general. Equivalently, a more detailed description of the survival process including tumour-immune interactions or immunotherapy can be used. Therefore, our framework is well suited to explore different model choices. These points may also be achieved by using agent-based models, they are in general computationally expensive and deriving a matching and complete set of decision rules can be challenging (Bull and Byrne 2023).

Even in the current form, our modelling framework provides a flexible starting point for exploring holistic models of cancer patients and enables future extensions that then accommodate more individualised stochastic processes and dependencies on various covariates such as therapy decisions. A modelling framework of this kind together with efficient inference is fundamentally important for studying cancer patient trajectories over time and providing practical value in clinical routine.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00285-025-02229-6.

Author Contributions V.W.: Conceptualization, formulation of the mathematical model, computation of the likelihood functions, implementation of simulation, implementation of parameter estimation and uncertainty quantification, preparation of figures, writing the original draft, review and editing. J.H.: Conceptualization, formulation of mathematical model, review and editing of the manuscript, funding acquisition.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2047-390685813, EXC 2151-390873048), by the German Federal Ministry of Education and Research (BMBF) under the CompLS program (GENImmune, Grant No. 031L0292F), by ERC grant INTEGRATE (Grant No. 101126146) and by the University of Bonn (via the Schlegel Professorship of J.H.).

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

Abrahamsson L, Humphreys K (2016) A statistical model of breast cancer tumour growth with estimation of screening sensitivity as a function of mammographic density. Stat Methods Med Res 25(4):1620–1637. https://doi.org/10.1177/0962280213492843

Abrahamsson L, Isheden G, Czene K, Humphreys K (2020) Continuous tumour growth models, lead time estimation and length bias in breast cancer screening studies. Stat Methods Med Res 29(2):374–395. https://doi.org/10.1177/0962280219832901
Agresti A (2013) Categorical data analysis. Wiley, Hoboken
Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705
Alzahrani E, Asiri A, El-Dessoky M, Kuang Y (2014) Quiescence as an explanation of Gompertzian tumor growth revisited. Math Biosci 254:76–82
Alzahrani EO, Kuang Y (2016) Nutrient limitations as an explanation of Gompertzian tumor growth. Disc Cont Dyn Syst B 21(2):357–372
Anderson DF (2007) A modified next reaction method for simulating chemical systems with time dependent propensities and delays. J Chem Phys 10(1063/1):2799998
Andrieu C, De Freitas N, Doucet A, Jordan MI (2003) An introduction to MCMC for machine learning. Mach Learn 50:5–43. https://doi.org/10.1023/A:1020281327116
Ankit R (2025) Webplotdigitizer. https://automeris.io
Arruebo M, Vilaboa N, Sáez-Gutierrez B, Lambea J, Tres A, Valladares M, González-Fernández Á (2011) Assessment of the evolution of cancer treatment therapies. Cancers 3:3279–3330. https://doi.org/10.3390/cancers3033279
Arvelo F, Sojo F, Cotte C (2016) Tumour progression and metastasis. Ecancermedicalscience. https://doi.org/10.3332/ecancer.2016.617
Ayuni Mazlan MS, Rosli N, Arief Ichwan SJ, Azmi NS (2017) Modelling the cancer growth process by stochastic differential equations with the effect of chondroitin sulfate (CS) as anticancer therapeutics. J Phys: Conf Ser 890:012085. https://doi.org/10.1088/1742-6596/890/1/012085
Baghestani AR, Moghaddam SS, Majd HA, Akbari ME, Nafissi N, Gohari K (2016) Survival analysis of patients with breast cancer using Weibull parametric model. Asian Pac J Cancer Prev 16(18):8567–8571
Bartoszynski R, Edler L, Hanin L, Kopp-Schneider A, Pavlova L, Tsodikov A, Zorin A, Yakovlev AY (2001) Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis. Math Biosci 171(2):113–142. https://doi.org/10.1016/S0025-5564(01)00058-X
Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: a fresh approach to numerical computing. SIAM Rev 59(1):65–98. https://doi.org/10.1137/141000671
Boire A, Burke K, Cox TR, Guise T, Jamal-Hanjani M, Janowitz T, Kaplan R, Lee R, Swanton C, Vander Heiden MG et al (2024) Why do patients with cancer die? Nat Rev Cancer 24(8):578–589
Box-Steffensmeier JM, Jones BS (2004) Event history modeling: a guide for social scientists. Cambridge University Press, Cambridge
Brito Fernandes O, Pais Silva Valente Cardoso Dias Carvalho, A, Klazinga N (2024) Beating cancer inequalities in the EU: spotlight on cancer prevention and early detection. Organisation for Economic Co-opertaion and Development. https://doi.org/10.1787/2074319x
Bull JA, Byrne HM (2022) The hallmarks of mathematical oncology. Proc IEEE. https://doi.org/10.1109/JPROC.2021.3136715
Bull JA, Byrne HM (2023) Quantification of spatial and phenotypic heterogeneity in an agent-based model of tumour-macrophage interactions. PLOS Comput Biol. https://doi.org/10.1371/journal.pcbi.1010994
Calvaresi D, Schumacher M, Calbimonte JP (2020) Agent-based modeling for ontology-driven analysis of patient trajectories. J Med Syst 44(9):158
Cancertrials.io (2022) Cancertrials.io. accessed on February 17, 2025; designed by the Laboratory of Systems Pharmacology at Harvard Medical School and Dana-Farber Cancer Institute and Palmer Laboratory at University of North Carolina https://labsyspharm.shinyapps.io/hmsclinical/,
Cheung LC, Albert PS, Das S, Cook RJ (2022) Multistate models for the natural history of cancer progression. Brit J Cancer 127(7), 1279–1288. https://doi.org/0.1038/s41416-022-01904-5
Clenshaw CW, Curtis AR (1960) A method for numerical integration on an automatic computer. Numer Math 2:197–205
Collins VP, Loeffler RK, Tivey H (1956) Observations on growth rates of human tumors. Am J Roentgenol Radium Therapy Nucl Med 76:988–1000
Cui W, Aouidate A, Wang S, Yu Q, Li Y, Yuan S (2020) Discovering anti-cancer drugs via computational methods. Front Pharmacol 11:733. https://doi.org/10.3389/fphar.2020.00733

Desmée S, Mentré F, Veyrat-Follet C, Sébastien B, Guedj J (2017) Nonlinear joint models for individual dynamic prediction of risk of death using Hamiltonian monte Carlo: application to metastatic prostate cancer. BMC Med Res Methodol 17:1–12. https://doi.org/10.1186/s12874-017-0382-9

d'Onofrio A, Fasano A, Monechi B (2011) A generalization of Gompertz law compatible with the Gyllenberg–Webb theory for tumour growth. Math Biosci 230:45–54

Elmore LW, Greer SF, Daniels EC, Saxe CC, Melner MH, Krawiec GM, Cance WG, Phelps WC (2021) Blueprint for cancer research: critical gaps and opportunities. CA Cancer J Clin 71(2):107–139. https://doi.org/10.3322/caac.21652

Engel J, Eckel R, Kerr J, Schmidt M, Fürstenberger G, Richter R, Sauer H, Senn HJ, Hölzel D (2003) The process of metastasisation for breast cancer. Eur J Cancer 39(12):1794–1806

Falzone L, Salomone S, Libra M (2018) Evolution of cancer pharmacological treatments at the turn of the third millennium. Front Pharmacol 9:421926. https://doi.org/10.3389/fphar.2018.01300

Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y (2020) Molecular principles of metastasis: a hallmark of cancer revisited. Signal Transduct Target Ther 5(1):28. https://doi.org/10.1038/s41392-020-0134-x

Franssen LC, Lorenzi T, Burgess AE, Chaplain MA (2019) A mathematical framework for modelling the metastatic spread of cancer. Bull Math Biol 81:1965–2010. https://doi.org/10.1007/s11538-019-00597-x

Gabbiani F, Cox SJ (2010) Stochastic Processes. In: Gabbiani F, Cox SJ (eds) Mathematics for neuroscientists. Elsevier, pp 251–266

Gasparini A, Humphreys K (2022) Estimating latent, dynamic processes of breast cancer tumour growth and distant metastatic spread from mammography screening data. Stat Methods Med Res 31(5):862–881. https://doi.org/10.1177/09622802211072496

Gerlee P (2013) The model muddle: in search of tumor growth laws. Can Res 73(8):2407–2411. https://doi.org/10.1158/0008-5472.CAN-12-4355

Goffe WL (1996) SIMANN: a global optimization algorithm using simulated annealing. Stud Nonlinear Dyn Econom. https://doi.org/10.2202/1558-3708.1020

Gompertz B (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of contingencies. Philos Trans R Soc 115:513–583. https://doi.org/10.1098/rspl.1815.0271

Greenspan HP (1972) Models for the growth of a solid tumor by diffusion. Stud Appl Math 51(4):317–340

Gyllenberg M, Webb GF (1990) A nonlinear structured population model of tumor growth with quiescence. J Math Biol 28(6):671–694

Hagman H, Frödin JE, Berglund Å, Sundberg J, Vestermark LW, Albertsson M, Fernebro E, Johnsson A (2016) A randomized study of KRAS-guided maintenance therapy with bevacizumab, erlotinib or metronomic capecitabine after first-line induction treatment of metastatic colorectal cancer: the nordic act2 trial. Ann Oncol 27(1):140–147

Hanahan D (2011) Hallmarks of cancer: new dimensions. Cancer Discov. https://doi.org/10.1158/2159-8290

Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100(1):57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hao SJ, Wan Y, Xia YQ, Zou X, Zheng SY (2018) Size-based separation methods of circulating tumor cells. Adv Drug Deliv Rev 125:3–20. https://doi.org/10.1016/j.addr.2018.01.002

Harshe I, Enderling H, Brady-Nicholls R (2023) Predicting patient-specific tumor dynamics: How many measurements are necessary? Cancers. https://doi.org/10.3390/cancers15051368

Heimann R, Hellman S (2000) Clinical progression of breast cancer malignant behavior: What to expect and when to expect it. J Clin Oncol 18(3):591. https://doi.org/10.1200/JCO.2000.18.3.591

Isheden G, Humphreys K (2019) Modelling breast cancer tumour growth for a stable disease population. Stat Methods Med Res 28(3):681–702. https://doi.org/10.1177/0962280217734583

Isheden G, Abrahamsson L, Andersson T, Czene K, Humphreys K (2019) Joint models of tumour size and lymph node spread for incident breast cancer cases in the presence of screening. Stat Methods Med Res 28(12):3822–3842. https://doi.org/10.1177/0962280218819568

Jakubowski W, Dobruch-Sobczak K, Migda B (2012) Errors and mistakes in breast ultrasound diagnostics. J Ultrasonogr 12(50):286. https://doi.org/10.15557/JoU.2012.0014

Johnson SG (2010) Notes on the convergence of trapezoidal-rule quadrature

Johnson SG (2020) Cubature.jl: The Cubature module for Julia. https://github.com/JuliaMath/Cubature.jl

Katsaounis D, Chaplain MAJ, Sfakianakis N (2023) Stochastic differential equation modelling of cancer cell migration and tissue invasion. J Math Biol 87(1):8. https://doi.org/10.1007/s00285-023-01934-4
Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680. https://doi.org/10.1126/science.220.4598.671
Klakattawi HS (2022) Survival analysis of cancer patients using a new extended Weibull distribution. PLoS ONE 17(2):e0264229
Koscielny S, Tubiana M, Le M, Valleron A, Mouriesse H, Contesso G, Sarrazin D (1984) Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination. Br J Cancer 49(6):709–715. https://doi.org/10.1038/bjc.1984.112
Kozusko F, Bajzer Ž (2003) Combining Gompertzian growth and cell population dynamics. Math Biosci 185(2):153–167
Kuang Y, Nagy JD, Eikenberry SE (2018) Introduction to mathematical oncology. Chapman and Hall/CRC, Boca Raton
L Rocha H, Aguilar B, Getz M, Shmulevich I, Macklin P (2024) A multiscale model of immune surveillance in micrometastases gives insights on cancer patient digital twins. npj Syst Biol Appl 10(1):144
Laird AK (1964) Dynamics of tumor growth. Br J Cancer 18:490–502. https://doi.org/10.1038/bjc.1964.55
Leader JJ (2022) Numerical analysis and scientific computation, 2nd edn. Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/978100304227
Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. Math Program 45:503–528. https://doi.org/10.1007/BF01589116
Mansour MBA, Abobakr AH (2022) Stochastic differential equation models for tumor population growth. Chaos Solitons Fractals 164:112738. https://doi.org/10.1016/j.chaos.2022.112738
Marušić M (1996) Mathematical models of tumor growth. Math Commun 1(2):175–192
Minn AJ, Gupta GP, Padua D, Bos P, Nguyen DX, Nuyten D, Kreike B, Zhang Y, Wang Y, Ishwaran H, Foekens JA, van de Vijver M, Massagué J (2007) Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci 104(16):6740–6745. https://doi.org/10.1073/pnas.0701138104
Mogensen PK, Riseth AN (2018) Optim: a mathematical optimization package for Julia. J Open Source Softw 3(24):615. https://doi.org/10.21105/joss.00615
Mustapha K, Gilli Q, Frayret JM, Lahrichi N, Karimi E (2016) Agent-based simulation patient model for colon and colorectal cancer care trajectory. Procedia Comput Sci 100:188–197
Nguyen Edalgo YT, Ford Versypt AN (2018) Mathematical modeling of metastatic cancer migration through a remodeling extracellular matrix. Processes 6(5):58. https://doi.org/10.3390/pr6050058
Norton L (1988) A Gompertzian model of human breast cancer growth. Can Res 48:7067–7071
Norton L (2005) Conceptual and practical implications of breast tissue geometry: toward a more effective, less toxic therapy. Oncologist 10(6):370–381. https://doi.org/10.1634/theoncologist.10-6-370
Ocaña-Tienda B, Eroles-Simó A, Pérez-Beteta J, Arana E, Pérez-García VM (2024) Growth dynamics of lung nodules: implications for classification in lung cancer screening. Cancer Imaging 24(1):113
Palmer S, Borget I, Friede T, Husereau D, Karnon J, Kearns B, Medin E, Peterse EF, Klijn SL, Verburg-Baltussen EJ et al (2023) A guide to selecting flexible survival models to inform economic evaluations of cancer immunotherapies. Value Health 26(2):185–192. https://doi.org/10.1016/j.jval.2022.07.009
Paquelet JR, Hendrick RE (2010) Lesion size inaccuracies in digital mammography. Am J Roentgenol 194:W115–W118. https://doi.org/10.2214/AJR.09.2927
Pavliotis GA (2014) Stochastic processes and applications. Texts in applied mathematics, vol 60. Springer, Berlin
Perera M, Dwivedi AK (2020) Statistical issues and methods in designing and analyzing survival studies. Cancer Rep 3(4):e1176. https://doi.org/10.1002/cnr2.1176
Plana D, Fell G, Alexander BM, Palmer AC, Sorger PK (2022) Cancer patient survival can be parametrized to improve trial precision and reveal time-dependent therapeutic effects. Nat Commun 13(1):873
Plevritis SK, Salzman P, Sigal BM, Glynn PW (2007) A natural history model of stage progression applied to breast cancer. Stat Med 26(3):581–595. https://doi.org/10.1002/sim.2550
Press W (2007) Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, Cambridge
Proust-Lima C, Taylor JM (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. Biostatistics 10(3):535–549. https://doi.org/10.1093/biostatistics/kxp009

Rackauckas C, Nie Q (2017) Differentialequations.jl—a performant and feature-rich ecosystem for solving differential equations in julia. J Open Res Softw 5(1):15. https://doi.org/10.5334/jors.151
Rahman MM, Islam MR, Rahman F, Rahaman MS, Khan MS, Abrar S, Ray TK, Uddin MB, Kali MSK, Dua K et al (2022) Emerging promise of computational techniques in anti-cancer research: at a glance. Bioengineering 9(8):335. https://doi.org/10.3390/bioengineering9080335
Revels J, Lubin M, Papamarkou T (2016) Forward-mode automatic differentiation in Julia. https://doi.org/10.48550/arXiv.1607.07892, arXiv: 2203.13043
Rizopoulos D (2011) Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 67(3):819–829. https://doi.org/10.1111/j.1541-0420.2010.01546.x
Ross KH, Gogineni K, Subhedar PD, Lin JY, McCullough LE (2019) Obesity and cancer treatment efficacy: existing challenges and opportunities. Cancer 125(10):1588–1592. https://doi.org/10.1002/cncr.31976
Rupp K, Lösch A, Hu YL, Nie C, Schill R, Klever M, Pfahler S, Grasedyck L, Wettig T, Beerenwinkel N, Spang R (2024) Modeling metastatic progression from cross-sectional cancer genomics data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btae250
Schwartz M (1961) A biomathematical approach to clinical tumor growth. Cancer 14(6):1272–1294
Schälte Y, Klinger E, Alamoudi E, Hasenauer J (2022) pyabc: efficient and robust easy-to-use approximate Bayesian computation. https://doi.org/10.48550/arXiv.2203.13043, arXiv:2203.13043
Schälte Y, Fröhlich F, Jost PJ, Vanhoefer J, Pathirana D, Stapor P, Lakrisenko P, Wang D, Raimúndez E, Merkt S, Schmiester L, Städter P, Grein S, Dudkin E, Doresic D, Weindl D, Hasenauer J (2023) pyPESTO: a modular and scalable tool for parameter estimation for dynamic models. Bioinformatics 39(11):btad711. https://doi.org/10.1093/bioinformatics/btad711
Sopik V, Narod SA (2018) The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer. Breast Cancer Res Treat 170:647–656. https://doi.org/10.1007/s10549-018-4796-9
Spratt JA, von Fournier D, Spratt JS, Weber EE (1993) Decelerating growth and human breast cancer. Cancer 71(6):2013–2019
Steel GG (1977) Growth kinetics of tumours. University Press, Oxford
Steel GG, Lamerton LF (1966) The growth rate of human tumours. Br J Cancer 20:74–86. https://doi.org/10.1038/bjc.1966.9
Strandberg R, Abrahamsson L, Isheden G, Humphreys K (2023) Tumour growth models of breast cancer for evaluating early detection-a summary and a simulation study. Cancers 15(3):912. https://doi.org/10.3390/cancers15030912
Tabassum S, Rosli NB, Binti Mazalan MSA (2019) Mathematical modeling of cancer growth process: a review. J Phys: Conf Ser 1366(1):012018. https://doi.org/10.1088/1742-6596/1366/1/012018
Talkington A, Durrett R (2015) Estimating tumor growth rates in vivo. Bull Math Biol 77(10):1934–1954. https://doi.org/10.1007/s11538-015-0110-8
Teschl G (2012) Ordinary differential equations and dynamical systems, vol 140. American Mathematical Society, Providence
Tjørve KMC, Tjørve E (2017) The use of Gompertz models in growth analyses, and new Gompertz-model approach: an addition to the unified-Richards family. PLoS ONE. https://doi.org/10.1371/journal.pone.0178691
Uhry Z, Hédelin G, Colonna M, Asselain B, Arveux P, Rogel A, Exbrayat C, Guldenfels C, Courtial I, Soler-Michel P, Molinié F, Eilstein D, Duffy S (2010) Multi-state Markov models in cancer screening evaluation: a brief review and case study. Stat Methods Med Res 19(5):463–486. https://doi.org/10.1177/0962280209359848
Vousden WD, Farr WM, Mandel I (2015) Dynamic temperature selection for parallel tempering in Markov chain monte Carlo simulations. Mon Notices R Astronom Soc. https://doi.org/10.1093/mnras/stv2422
Wolfram Research, Inc. (2024) Mathematica, Version 12.0. https://www.wolfram.com/mathematica, Champaign, IL
Wolin KY, Carson K, Colditz GA (2010) Obesity and cancer. Oncologist 15(6):556–565. https://doi.org/10.1634/theoncologist.2009-0285
World Health Organization (2018) Cancer. https://www.who.int/health-topics/cancer
World Health Organization (2024) Obesity. https://www.who.int/health-topics/obesity
Wright S (1926) Book review on: Raymond Pearl and J. Shirley Sweeney "The Biology of Population Growth". J Am Stat Assoc 21(156):493–497
Wu L, Liu W, Yi GY, Huang Y et al (2012) Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. J Probab Stat. https://doi.org/10.1155/2012/640153

Yin A, Moes DJA, van Hasselt JG, Swen JJ, Guchelaar HJ (2019) A review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors. CPT: Pharmacom Syst Pharmacol 8(10):720–737. https://doi.org/10.1002/psp4.12450

Zhang L, Mosquera I, Lucas E, Rol ML, Carvalho AL, Basu P (2023) Canscreen5, a global repository for breast, cervical and colorectal cancer screening programs. Nat Med 29(5):1135–1145. https://doi.org/10.1038/s41591-023-02315-6