Society for
Mathematical
Biology

Check for
updates

Predicting Radiotherapy Patient Outcomes with Real-Time Clinical Data Using Mathematical Modelling

Alexander P. Browning[1] · Thomas D. Lewin[1,2] · Ruth E. Baker[1] ·
Philip K. Maini[1] · Eduardo G. Moros[3] · Jimmy Caudell[3] · Helen M. Byrne[1] ·
Heiko Enderling[3,4,5]

Abstract

Longitudinal tumour volume data from head-and-neck cancer patients show that tumours of comparable pre-treatment size and stage may respond very differently to the same radiotherapy fractionation protocol. Mathematical models are often proposed to predict treatment outcome in this context, and have the potential to guide clinical decision-making and inform personalised fractionation protocols. Hindering effective use of models in this context is the sparsity of clinical measurements juxtaposed with the model complexity required to produce the full range of possible patient responses. In this work, we present a compartment model of tumour volume and tumour composition, which, despite relative simplicity, is capable of producing a wide range of patient responses. We then develop novel statistical methodology and leverage a cohort of existing clinical data to produce a predictive model of both tumour volume progression and the associated level of uncertainty that evolves throughout a patient's course of treatment. To capture inter-patient variability, all model parameters

✉ Alexander P. Browning
browning@maths.ox.ac.uk

✉ Heiko Enderling
HEnderling@mdanderson.org

1 Mathematical Institute, University of Oxford, Oxford, UK

2 Roche Pharma Research and Early Development, Roche Innovation Center, Basel, Switzerland

3 Department of Radiation Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, USA

4 Department of Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, USA

5 Present Address: Department of Radiation Oncology, MD Anderson Cancer Center, Houston, TX, USA

are patient specific, with a bootstrap particle filter-like Bayesian approach developed to model a set of training data as prior knowledge. We validate our approach against a subset of unseen data, and demonstrate both the predictive ability of our trained model and its limitations.

Keywords Head-and-neck cancer · Predictive model · Patient variability · Heterogeneity · Uncertainty · Radiotherapy

1 Introduction

Radiotherapy remains a mainstay of cancer treatment, with approximately half of all cancer patients receiving radiotherapy as part of their standard of care (Fowler 2006; Torres-Roca 2012; Enderling et al. 2009). It is common for a patient's course of treatment to be determined solely by tumour etiology, location, and stage. Other patient-specific factors, such as the intrinsic radiosensitivity and composition of a tumour, are not typically used to inform protocol selection in the clinic (Caudell et al. 2017). Clinical studies suggest that patients at a similar tumour, node, and metastasis (TNM) stage, and with comparable pre-treatment tumour volumes, may respond differently to the same radiotherapy fractionation schedule (Scott et al. 2017; Sunassee et al. 2019). Mathematical models have the potential to capitalise on real-time clinical observations to both predict patient specific responses and guide clinical decision-making. It is hoped that such a tight integration could eventually be used to personalise fractionation schedules either *a priori* or adaptively during a patient's course of treatment (Enderling et al. 2019).

Challenges associated with the application of mathematical models to interpret data and draw predictions are perhaps most acute for single-patient clinical data. Models must be sufficiently complex to reproduce the full gamut of patient responses (Yankeelov et al. 2013; Collis et al. 2017; Brady and Enderling 2019). However, clinical data are often limited, typically comprising solely noisy measurements of the gross tumour volume (GTV) at sparse time intervals throughout a patient's course of treatment (Brady and Enderling 2019; Harshe et al. 2023). The necessity to start treatment as soon as possible after diagnosis means that pre-treatment predictions are often drawn from only one or two observations. Consequently, models aimed at clinical application are relatively simple (Prokopiou et al. 2015; Rockne and Frankel 2017), incorporate limited biological detail, and often describe only the time-evolution of the GTV (Sunassee et al. 2019; Prokopiou et al. 2015; Rockne and Frankel 2017). While simplicity can elicit parameter identifiability and avoid overfitting, predictions can be poor—or even misleading—if a model is so simple as to be unable to capture the range of observed (possible) responses. The dangers of overfitting are particularly pronounced for single-patient clinical data used for prediction, where model validation must be assessed pre-treatment; in diametric opposition to experimental data, technical replicates are never available. It is, therefore, crucial to validate models across a wide range of responses, and to accurately quantify uncertainty in predictions used in clinical decision-making (Brady and Enderling 2019).
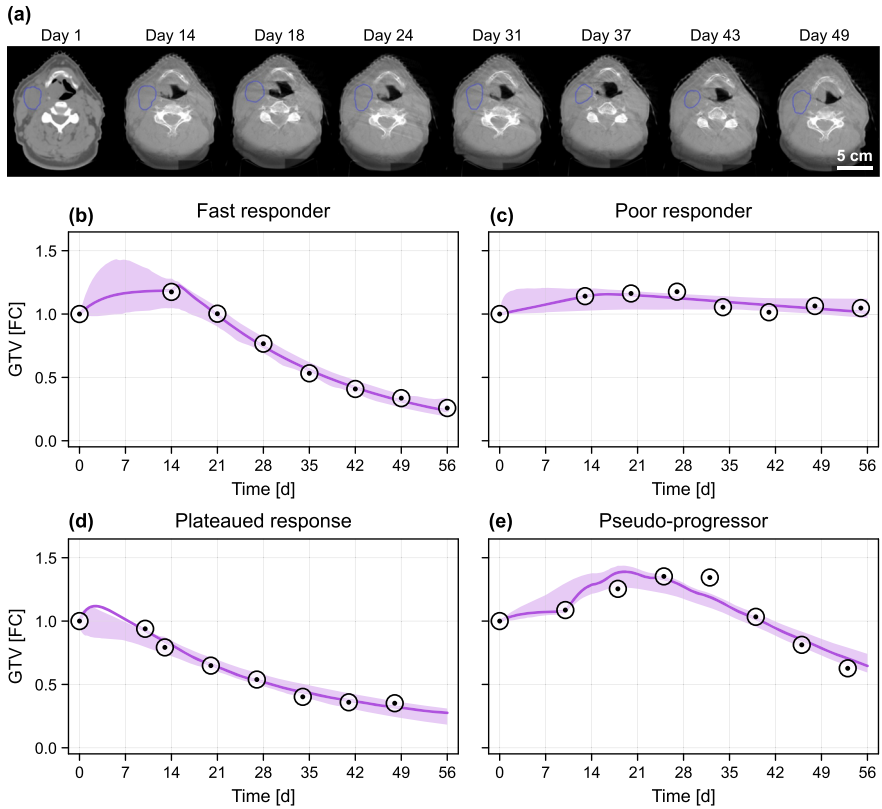
Fig. 1 Gross tumour volume (GTV) measurements taken during radiotherapy. **a** Example CT scan of an oropharyngeal cancer patient throughout treatment, showing tumour contoured in blue (data courtesy of CD Fuller and MD Anderson). **b–e** Clinical data representing four qualitatively different radiotherapy results. Predictions from the mathematical model, along with a 95% credible interval for the modelled observation means, are shown in purple. In all cases, the radiotherapy schedule starts at the time of the second observation. The four patients shown are excluded from the training data analysed in later parts of our study. The units of GTV are given as the fold change (FC) relative to the initial volume (colour figure online)

In this work, we present a predictive mathematical modelling framework using clinical GTV data from a previously published cohort of head-and-neck cancer patients who exhibit a variety of treatment responses (Fig. 1) (Zahid et al. 2021a, b). The primary goal of our framework is to integrate previously observed clinical observations to predict the time course of radiotherapy response in new patients. To demonstrate our framework, we focus our analysis on prediction of the tumour volume progression in four patients presented in Fig. 1 and in our previous work Lewin et al. (2020): these patients are excluded from the otherwise randomly-selected cohort of patients used to train the mathematical model. All patients in the clinical data set receive a standard radiotherapy fractionation schedule, comprising fractions of 2 Gy delivered on weekdays over a four- to seven-week period (Lewin et al. 2016). To keep our study as widely applicable as possible, we work with the most fundamental, albeit limited,

mode of single patient data. Computed tomography (CT) scans are routinely used to image tumours pre-treatment at both the diagnosis and treatment planning stages (Fig. 1a) (Stevens et al. 2013; Sharma et al. 2016; Wang et al. 2009). Further scans, such as cone beam CT, may be taken upon the delivery of each fraction but are usually used solely for alignment purposes; for our data, scans were available once per week during treatment. These CT scans are not of a high spatial resolution, are noisy, of a low contrast, and do not differentiate heterogeneity in tumour composition. As such, only noisy measurements related to an estimate of the GTV are available at relatively sparse intervals throughout each patient's course of treatment (typically, once per week). The heterogeneity in radiotherapy response exhibited in Fig. 1b–e raises several important questions: in particular, how early into treatment can a practitioner determine if a patient is responsive, and to what extent is it possible to predict the final tumour volume during treatment using only GTV measurements? Given the side-effects associated with radiotherapy, and possible indirect costs of switching treatments at too late a TNM stage, any improvement in prediction accuracy is of great clinical value.

Mathematical models of tumour progression vary significantly in complexity; ranging from simple phenomenological models of GTV, such as logistic and Gompertz growth (Sachs et al. 2001; McAneney and O'Rourke 2007; Rockne et al. 2010; Chvetsov 2013; Prokopiou et al. 2015; Tariq et al. 2016; Poleszczuk et al. 2018; Browning and Simpson 2023), to highly detailed spatial models that capture multiple facets of tumour heterogeneity (Greenspan 1972; Rockne et al. 2010; Lewin et al. 2018, 2020; Browning and Simpson 2023). The limitations and challenges imposed by clinical data yield an overrepresentation of the former, meaning that the functional forms for both growth and radiotherapy response are motivated almost entirely by empirical observations rather than the underlying biological mechanisms. Yet, it is now well established that intra-tumour heterogeneity and the tumour microenvironment play important roles in overall growth, and may significantly influence treatment outcome (Ribba et al. 2006; Rockne et al. 2009, 2015; Lewin et al. 2018, 2020; Browning et al. 2021). Motivated by these findings and in consideration of the noisy data available for prediction, we take an intermediate approach and utilise a two compartment extension of the so-called proliferation-saturation-index (PSI) model of Prokopiou et al. (2015) and later Poleszczuk et al. (2018). This choice of ordinary differential equation (ODE) model balances simplicity, through a phenomenological description of radiation-free tumour growth saturation, with biological detail, through a radiotherapy response corresponding to a transfer of cellular material from a living to a dead state. Compared with purely statistical or machine learning models, our mathematical approach allows a full, interpretable, integration of clinical data from individuals, whereby the radiotherapy schedule is imported directly from the reported patient fractionation schedule. Finally, our model contains sufficient detail to allow us to quantify the potential utility of expanding clinical data collection to include information relating to tumour composition in addition to GTV.

We take a Bayesian pseudo-hierarchical approach to inference and model calibration, by leveraging observed population-level information to draw predictions and quantify corresponding levels of prediction uncertainty. To account for inter-patient heterogeneity, all model parameters are allowed to vary between patients. A schematic of the approach is provided in Fig. 2. In contrast to standard Bayesian hierarchical
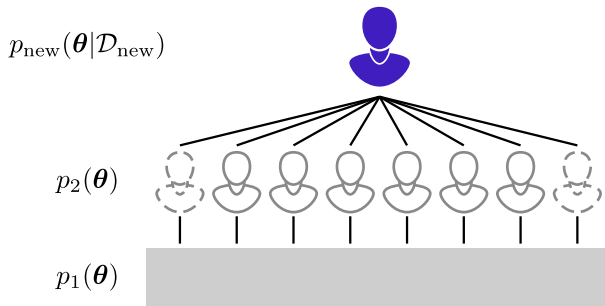
Fig. 2 Pseudo-hierarchical approach used in the analysis. Devoid of any data, knowledge about model parameters is encoded in the "first-level prior", denoted by $p_1(\boldsymbol{\theta})$, and is used to individually form a set of posterior distributions for patients in the training set. The "second-level prior", denoted by $p_2(\boldsymbol{\theta})$, represents knowledge gained from analysis of the training set and is used to form posterior distributions for *new* patients, denoted by $p_{\text{new}}(\boldsymbol{\theta}|\mathcal{D}_{\text{new}})$. In effect, the approach identifies patients in the training set with possibly similar outcomes to new patients

approaches, we do not make any parametric assumptions relating to the distribution of model parameters between individuals. Instead, we build up a population-level posterior distribution by calibrating the model to individual patients in a set of training data. As illustrated in Fig. 2, this population-level posterior distribution forms the prior for analysis of new patients. In effect, the prior distribution for new patients quantifies a set of possible patient outcomes based on those observed in the training data. Successively applying the Bayesian inference algorithm as data becomes available throughout a patient's course of treatment allows us to update this potential set of future outcomes and the corresponding uncertainty in tumour volume. We validate our approach by first exploring prediction ability on synthetic data, and then prospectively making predictions on the four patients presented in Fig. 1b–e as they undergo their course of treatment.

2 Methods

In this section, we outline the clinical data, and the mathematical and statistical methodology developed and later employed in this work. First, in Sect. 2.1 we describe and present the clinical data set used for quantitative analysis and which demonstrate four disparate treatment response classifications. Secondly, in Sect. 2.2 we present a mechanistic mathematical model of tumour volume progression, along with a set of objective criteria that we use to classify model realisations into the four observed classifications. Subsequently, in Sect. 2.3 we present a statistical model that connects model predictions to clinical measurements. In Sect. 2.4 we outline the novel statistical methodology employed in the analysis. Finally, in Sect. 2.5 we outline the procedure for resampling from the joint posterior to produce synthetic patient data. A Julia implementation of the model and inference algorithm, along with data used in the analysis, are available on GitHub.[1]

---

[1] https://github.com/ap-browning/clinical-inference

2.1 Tumour Volume Data

Current clinical practice involves two CT scans collected for each patient; one at diagnosis and one at treatment planning. These scans are then used to estimate GTV (Wang et al. 2009; Stevens et al. 2013; Sharma et al. 2016). While it is feasible to obtain further scans at the time of delivery of each fraction, these scans are often of a low quality, being used primarily to position the patient. As such, they are not typically stored for research purposes.

In this paper, we use retrospective volumetric data, collected weekly, from head-and-neck cancer patients, across multiple anatomical locations, including the oropharynx, tonsil and base of tongue. Patients were immobilised via a thermoplastic mask with or without bite block. Isocenter and positioning was verified daily via orthoganal kV or CBCT imaging. Each CT scan was segmented by the same radiation oncologist, giving weekly tumour volumes throughout treatment in addition to a volume measurement at the treatment planning stage. Weekly cone beam CT (CBCT) scans were extracted from the record and verify system (Mosaiq, Elekta). Suitable CBCTs with minimal artifact were selected for contouring. Clinical target volume (CTV) was created from GTV with a 5 mm isotropic expansion. CTV was then trimmed from barriers to spread including air, bone, fascial planes, and in some cases muscle. Planning target volume (PTV) was created from CTV via 3 mm isotropic expansion. An example suite of contoured CT scans from a single patient is shown in Fig. 1a. The GTV data shown in Fig. 1b–e correspond to those presented and discussed in Lewin et al. (2020). In total, GTV data from 51 patients was collected and made available as supplementary material. All methods were carried out in accordance with the institutional policies of the Moffitt Cancer Center. The clinical protocol covering patient data and methods used in this paper was approved by the Moffitt Cancer Center's Institutional Review Board (IRB). Since this is a retrospective study using de-identified data of adult human subjects, informed consent was waived by the IRB.

2.2 Mathematical Model

Mathematical models of tumour growth and, to a lesser extent, radiotherapy response, are well established (Araujo and McElwain 2003), ranging from compartmental ODE models (Sachs et al. 2001; Chvetsov et al. 2009; Wang and Feng 2013; Chvetsov et al. 2014; Prokopiou et al. 2015; Tariq et al. 2016; Sunassee et al. 2019; Browning and Simpson 2023) and spatially-resolved partial differential equation models (Greenspan 1972; Rockne et al. 2009, 2010, 2015; Lewin et al. 2018; Browning et al. 2021; Browning and Simpson 2023), to agent-based models (Enderling et al. 2009; Gao et al. 2013; Alfonso et al. 2014; Powathil et al. 2013, 2016; Richard et al. 2007) and purely probabilistic models (Zaider and Minerbo 2000; Hanin 2004; Gong et al. 2011; Bobadilla et al. 2017).

Given the limitations imposed by GTV clinical data, we present a relatively simple mathematical model that is able to capture the four classes of tumour response observed in Fig. 1. In particular, we extend the PSI model (Poleszczuk et al. 2018) to include a simple measure of tumour composition by modelling the volume of both living

cells, $L(t)$, and necrotic debris, $N(t)$. The GTV is given by $V(t) = L(t) + N(t)$. Living cells proliferate logistically with rate $\lambda \, \mathrm{d}^{-1}$ and carrying capacity $K \, [V(t)]$, and potentially undergo necrosis at rate $\eta \, \mathrm{d}^{-1}$. Given that the growth dynamics occur on a much slower timescale than the interval during which the patient receives each fraction, we model radiotherapy as an instanaeous transfer of living cells to necrotic debris at record-informed dosing times t_i, $i = 1, 2, \ldots, n$. The model equations are given by

$$
\frac{dL}{dt} = \overbrace{\lambda L \left(1 - \frac{L}{K} \right)}^{\text{Growth}} - \overbrace{\eta L}^{\text{Necrosis}} - \overbrace{\gamma L \sum_{i=1}^{n} \delta(t - t_i)}^{\text{Radiotherapy}},
$$

$$
\frac{dN}{dt} = \eta L - \underbrace{\zeta N}_{\text{Decay}} + \gamma L \sum_{i=1}^{n} \delta(t - t_i),
$$

(1)

where $\delta(t - t_i)$ is a delta function, representing a transfer of a volume γL from the living compartment to the dead compartment, such that $\gamma \, \mathrm{d}^{-1}$ quantifies the strength of radiotherapy response. We assume further that necrotic material is degraded at a constant rate $\zeta \, \mathrm{d}^{-1}$. To capture inter-patient heterogeneity, all parameters are allowed to vary between patients (Lawson et al. 2018).

The data suggest that initial GTV is comparable between responsive and poorly responsive patients (Table 1). Therefore, we normalise $L(t)$ and $N(t)$ with the initial GTV such that $V(0) = 1$ and describe the initial tumour composition as

$$
L(0) = 1 - \phi_0, \qquad N(0) = \phi_0,
$$

(2)

where $0 \le \phi_0 \le 1$ is an unknown, patient-specific parameter to be estimated that represents the proportion of the tumour occupied by dead material at $t = 0$. We note further that the interpretation of the carrying capacity parameter K is with respect to the measured initial GTV. Thus, GTV measurements presented throughout the paper may be interpreted as the fold change (FC) compared to the initial GTV. The interpretation of all other parameters remains unchanged by this choice of units.

In the supplementary material (Figs. S1 and S2), we perform a parameter sweep across parameters relating to necrosis and necrotic material decay (η and ζ, respectively), for a patient subject to daily doses of radiotherapy on weekdays over a six week period, to verify that the model is able to reproduce the wide range of dynamics observed in the clinical data. While the parameter sweep is not exhaustive, the results demonstrate that varying only these two parameters is sufficient to produce the range of responses observed in Fig. 1.

2.2.1 Classifying Responses

We observe four classes of qualitative response within the clinical data, as highlighted in Fig. 1 and summarised in Table 1. In Fig. 1b, the patient responds well to radiotherapy, with the tumour decreasing markedly in volume throughout treatment. Hereafter, we refer to a patient exhibiting this type of behaviour as a *fast responder*. By contrast,

Table 1 Prior classification of each patient response class, based on the full posterior, $p(\boldsymbol{\theta}|\{\mathcal{D}_i\}_{i=1}^n)$ and the second-level prior $p_2(\boldsymbol{\theta})$, the latter corresponding to an expanded kernel density estimate constructed from samples of the full posterior

Classification	Proportion $p(\boldsymbol{\theta}|\{\mathcal{D}_i\}_{i=1}^n)$	$p_2(\boldsymbol{\theta})$	Initial volume [cm^3] Mean	Std.	Count
(∗) Fast responder	0.8763	0.6278	16.8	11.6	35.0
Poor responder	0.0598	0.3473	20.2	7.3	2.4
(∗) Plateaued response	0.0035	0.0021	3.4	4.8	0.1
(∗) Pseudo-progression	0.0604	0.0228	13.7	12.0	2.4
Eventual response (∗)	0.9402	0.6527	16.6	11.7	37.6

The statistics related to the initial volume are based on the classifications of the prior samples corresponding to each patient in the training set, hence non-integer counts arise due to probabilistic classification of patients. An approximate statistical test, based on Welch's approximate unequal variance t-test (Welch 1947), indicates no statistically significant difference between fast and poor responders ($P = 0.582$), nor between responders (∗) and poor responders ($P = 0.557$). Asterisks indicate classifications corresponding to patients who show an eventual response

there are patients for whom the effects of radiotherapy appear to be marginal when viewed in terms of tumour volume over time alone, as is the case in Fig. 1c. We classify these patients as *poor responders*. In a number of cases, the initial response of the tumour to radiotherapy appears to be favourable, but the response plateaus in the latter stages of treatment, resulting in a non-negligible final tumour volume (Fig. 1d). Such patients are classified as having a *plateaued response*. However, this radiographic volume may subsequently recede in the weeks after radiotherapy. Occasionally, as in Fig. 1e, a patient may appear to exhibit continued tumour progression throughout the first few weeks of radiotherapy before showing a delayed response, characterised by a decrease in tumour volume towards the end of treatment. We characterise this type of response as *pseudo-progression*.

We classify a model realisation into one of four classes of response based on a standard patient receiving doses on weekdays over a six week period, with CT measurements taken at the start of each treatment week and at the time of the final dose (the pre-treatment volume measurement is not used to classify patients). Based on the set of noise free synthetic measurements generated from the model, we define each classification according to the following quantitative criteria.

1. *Poor responder.* All measurements above 85% of the volume observed at the start of treatment.
2. *Responder.* At least one measurement below 85% of the volume observed at the start of treatment. Responders are further classified:

   (a) *Pseudo-progressor.* A second (noise-free) measurement greater than 102% of the first following radiotherapy onset.
   (b) *Plateaued response.* Not a pseudo-progressor, with a final measurement greater than 20% of the initial, and with a final rate-of-change less than 10% of the maximum rate-of-change observed.
   (c) *Fast responder.* Not in any other classification.

The specific thresholds chosen in the classification algorithm yield excellent results that reliably distinguish between each class (Fig. S4). However, the relatively small number of plateaued responders and pseudo-progressors in the training set (Table 1) suggests that the criteria will need to be reassessed should more data become available.

2.3 Statistical Model

We take a standard approach and assume that CT scan data are independent and normally distributed about the model prediction (Kreutz et al. 2012) such that

$$V_{\text{obs}} \sim \mathcal{N}\left(V_{\text{total}}, \sigma^2(V_{\text{total}})\right), \tag{3}$$

where the standard deviation

$$\sigma(V_{\text{total}}) = \alpha_1 + \alpha_2 V_{\text{total}}, \tag{4}$$

is assumed to be a linear function of GTV such that the statistical model captures both additive and multiplicative normal noise: α_1 represents an absolute contribution to the variance, and α_2 a relative contribution.

While the dynamical parameters are assumed to vary between patients, we assume that the noise parameters remain fixed. Therefore, we pre-estimate the noise parameters α_1 and α_2 by first inferring them alongside dynamical parameters for each patient. We then pool an equal number of noise parameter posterior samples for each patient and approximate (α_1, α_2) as the marginal posterior mode. We are motivated to take this relatively standard approach of pre-estimating the noise parameters to reduce both the dimensionality of the parameter space and the complexity of the statistical methodology.

2.4 Bayesian Inference

An important difference between clinical and experimental data relates to the sample size: in clinical studies, each patient undergoes therapy only once. Given that patients are highly heterogeneous and data are relatively limited (Fig. 1), this poses a significant statistical challenge for computational inference. To account for this, we take a pseudo-hierarchical approach to inference and prediction by first training the model on a subset of the data (the training set). We are motivated to develop this novel approach to inference as opposed to a more standard Bayesian hierarchical approach as there is no sensible means by which to propose a particular distributional form for the joint parameter distributions at the population-level: given the distinct classes of response observed in Fig. 1, for example, we expect the joint parameter distribution to be multimodal. The correlation structure between model parameters is also unclear.

From a full cohort of 51 patients, we randomly select a group of 40 patients to act as the *training* set; these patients represent those that have been observed throughout an entire course of treatment, prior to the present. For each patient in the training set,

Table 2 Parameters and first-level prior distributions

Parameter	Units	Prior	Description
λ	d^{-1}	$\log \lambda \sim \mathcal{U}(-10, 0)$	Cell proliferation rate
K	–	$\log K \sim \mathcal{U}(0, 5)$	Carrying capacity
γ	d^{-1}	$\log \gamma \sim \mathcal{U}(-10, 0)$	Radiotherapy response
ζ	d^{-1}	$\log \zeta \sim \mathcal{U}(-10, 3)$	Necrotic debris decay rate
η	d^{-1}	$\log \eta \sim \mathcal{U}(-10, 3)$	Cell necrosis rate
ϕ_0	–	$\log \phi_0 \sim \mathcal{U}(-5, 0)$	Initial necrotic proportion

The description relates to the exponentiated log parameter

we assume that initial knowledge about the model parameters is encoded in a "first-level prior", $p_1(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\log \lambda, \log K, \log \gamma, \log \zeta, \log \eta, \log \phi_0)$ (Fig. 2). We then update our knowledge about the parameters pertaining to patient i using Bayes theorem such that

$$\underbrace{p^{(i)}(\boldsymbol{\theta}|\mathcal{D}_i)}_{\text{Posterior } i} \propto \underbrace{p(\mathcal{D}_i|\boldsymbol{\theta})}_{\text{Likelihood}} p_1(\boldsymbol{\theta}), \tag{5}$$

where $\mathcal{D}_i$ represents data (including both volume measurements and the radiotherapy schedule) for patient i. We choose $p_1(\boldsymbol{\theta})$ to an independent multivariate uniform (see Table 2 and Fig. 3), an uninformative choice.

The posterior for patient i can be interpreted as the full posterior, conditioned on knowledge that the parameters relate to patient i

$$p^{(i)}(\boldsymbol{\theta}|\mathcal{D}_i) = p(\boldsymbol{\theta}|\{\mathcal{D}_i\}_{i=1}^n, i). \tag{6}$$

The *full posterior* can be obtained by marginalising over all patients in the training set and is given by

$$p(\boldsymbol{\theta}|\{\mathcal{D}_i\}_{i=1}^n) = \sum_i w_i \, p^{(i)}(\boldsymbol{\theta}|\mathcal{D}_i), \tag{7}$$

where $w_i = \mathbb{P}(i)$ represents the prior probability (i.e., weighting) of patient i. The result in Eq. (7) follows immediately from Eq. (6) by the law of total probability. For simplicity, we set $w_i = $ const, however, such weights may be allowed to differ if additional knowledge informs patient similarity; for example, based on characteristics known to affect radiotherapy response, such as the clinical stage or age of a patient (Belgioia et al. 2021). Another way to interpret the full posterior is that of a uniform mixture of the individual-level posterior distributions. We then denote the full posterior as the "second-level prior", $p_2(\boldsymbol{\theta})$, which represents our knowledge about the parameters when analysing *new* patients (we drop notational dependence on already observed data for convenience) (Fig. 2). An interpretation of our procedure is to identify the similarity between the new patient and the observed treatment outcomes for patients in the training set, and to combine the additional knowledge obtained from past patients when predicting outcomes for the new patient. In Fig. 3 we compare the first-level prior $p_1(\boldsymbol{\theta})$ to the full posterior (Eq. (7)), and in Fig. 4 we show pairwise marginal distributions of samples from the full posterior (Eq. (7)).
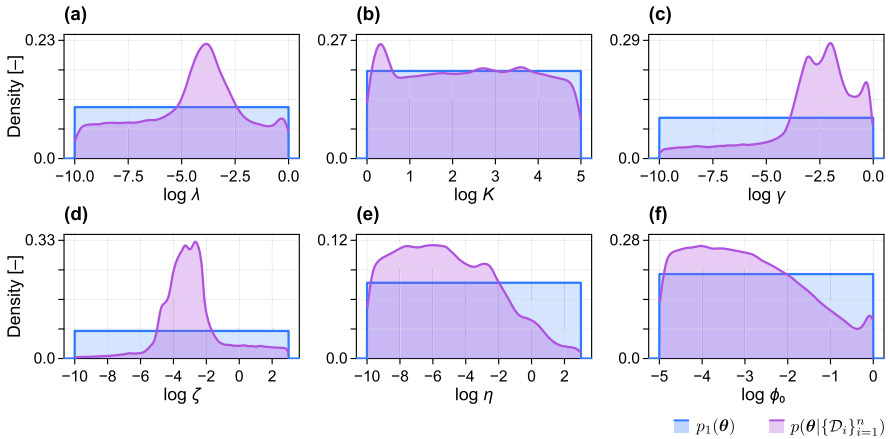
Fig. 3 Parameter posteriors from analysis of training data. First-level prior distribution (blue) and full posterior (purple) following analysis of the training data. The first-level prior, $p_1(\boldsymbol{\theta})$, comprises independent uniform distributions in the log of each unknown parameter. Parameters relate to the cell proliferation rate, λ, the carrying capacity, K, the radiotherapy response strength, γ, the decay rate of necrotic debris, ζ, the cell necrosis rate, η, and the initial proportion of the population that is necrotic, ϕ_0 (colour figure online)

Given a possibly temporally incomplete set of measurements from a new patient, $\mathcal{D}_{\text{new}}$, the posterior distribution of the parameters is again given by

$$p_{\text{new}}(\boldsymbol{\theta}|\mathcal{D}_{\text{new}}) \propto p(\mathcal{D}_{\text{new}}|\boldsymbol{\theta})\, p_2(\boldsymbol{\theta}). \qquad (8)$$

A simple technique to obtain a set of weighted samples from $p_{\text{new}}(\boldsymbol{\theta}|\mathcal{D}_{\text{new}})$ is to apply a bootstrap particle filter to pre-obtained samples from $p_2(\boldsymbol{\theta})$. Since patients in the training set are weighted equally, these may comprise a concatenation of samples from each posterior (we obtain these using an adaptive MCMC algorithm (Vihola 2020), diagnostic statistics and convergence plots are given as supplementary material). An advantage of the bootstrap particle filter approach is that it requires minimal computational effort to update the posterior for new patients. The primary limitation introduced by this choice is that we cannot distinguish between parameters that vary between patients and those that are fixed: hence, we pre-estimate and fix the noise parameters in this work.

In practice, this approach may be problematic since patients in the training set are unlikely to be identically representative of new patients, particularly for small training sets (in our case, $n = 40$). In the bootstrap particle filter, this would lead to a small number of heavily weighted particles (that may or may not produce model realisations similar to the new patient data). We address this potential issue by forming $p_2(\boldsymbol{\theta})$ by resampling perturbed particles from $p(\boldsymbol{\theta}|\{\mathcal{D}_i\}_{i=1}^n)$ using a multivariate normal distribution with covariance matrix, denoted Σ_ε, constructed by expanding the covariance matrix of Silverman's rule for kernel density estimation,

$$\Sigma_\varepsilon = \beta \left(\frac{4}{m(\dim(\boldsymbol{\theta}) + 2)} \right)^{\frac{1}{\dim(\boldsymbol{\theta})+4}} \operatorname{diag}\left(\Sigma_\theta\right), \qquad (9)$$

where β is an expansion factor (we choose $\beta = 2$), m is the number of samples of $\boldsymbol{\theta}|\{\mathcal{D}_i\}_{i=1}^n$ and $\Sigma_{\boldsymbol{\theta}}$ is the covariance matrix of the samples. We reject samples outside the support of the first-level prior $p_1(\boldsymbol{\theta})$ (see Table 2), in effect constructing $p_2(\boldsymbol{\theta})$ as a kernel density estimate with truncated multivariate normal kernels. This approach is also similar to a one-step sequential Monte Carlo algorithm (Moral et al. 2006).

2.4.1 Quantifying Goodness-of-Fit

We quantify goodness-of-fit using the so-called Bayesian R^2 statistic (Gelman et al. 2019), defined for a single posterior sample by

$$R^2 = \frac{\text{Var}(V_{\text{fit}})}{\text{Var}(V_{\text{fit}}) + \text{Var}(V_{\text{fit}} - V_{\text{obs}})}, \tag{10}$$

where V_{fit} denotes the set of fitted values, and V_{obs} denotes the set of observed values. A given posterior distribution yields a distribution of R^2 statistics: in this work, we report the median of the resultant distribution. Similarly to the frequentist R^2 statistic, a Bayesian R^2 statistic of unity indicates that the model captures all data variability (i.e., the variance of residuals, $\text{Var}(V_{\text{fit}} - V_{\text{obs}})$, is zero), while a Bayesian R^2 statistic of zero indicates that all fitted values lie on a horizontal line (hence, we expect low R^2 statistics for poor responders).

2.5 Generation of Synthetic Patient Data

We generate synthetic patient data by resampling parameters from the full posterior and exposing patients to what we have previously referred to as a standard radiotherapy regime (weekday doses over a six week period, with CT measurements taken at the start of each treatment week and at the time of the final dose). Noise is added to synthetic measurements according to the statistical model (Sect. 2.3) with pre-estimated noise parameters. Synthetic data from a patient exhibiting a specific classification are produced by utilizing only full posterior samples that produce the classification of interest.

3 Results and Discussion

3.1 Model Calibration and Patient Classification

To verify that the two compartment model can capture the range of radiotherapy responses observed *in situ*, we first calibrate the untrained mathematical model to data from single patients in Fig. 1b–e using MCMC with the first-level prior. Best fits, along with the associated uncertainty in GTV, are shown alongside data in Fig. 1b–e. Overall, the model is able to reproduce clinical observations, although it has some difficulty distinguishing between fast responders and plateaued responses. Given that the plateaued response in Fig. 1d is diagnosed as such from only the last three observations, we attribute the potential for misclassification to uncertainty in the clinical
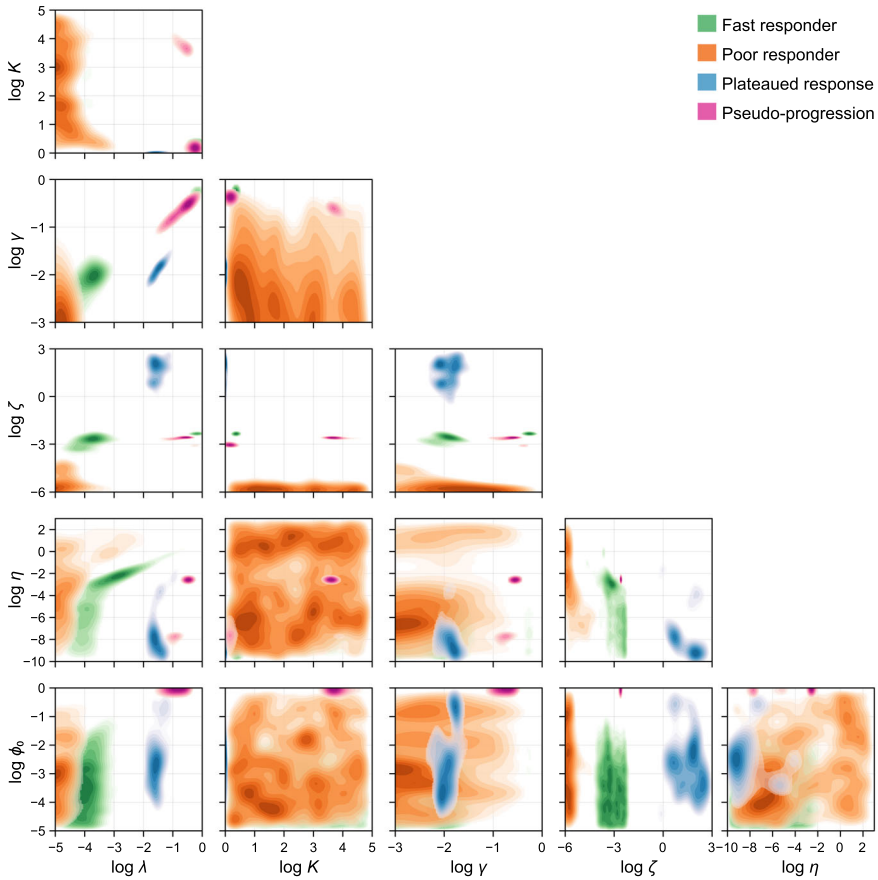
Fig. 4 Parameter clustering according to classified patient response. Kernel density of the full posterior distribution, following analysis of the training data set. Samples are classified into one of four patient responses according to criteria set out in Sect. 2.2.1, and kernel density estimates of bivariate marginal distributions conditioned on each classification shown. To aid comparison in the vicinity of the mode of each conditional posterior, only regions with densities greater than 50% of the maximum are shown. Parameters relate to the cell proliferation rate, λ, the carrying capacity, K, the radiotherapy response strength, γ, the decay rate of necrotic debris, ζ, the cell necrosis rate, η, and the initial proportion of the population that is necrotic, ϕ_0

observations (i.e., the noise model) and the lack of proceeding data points; it is impossible to tell whether this patient will continue to respond should treatment continue. Similar results are also seen for synthetic patients in the supplementary material, where patients that actually exhibit a plateaued response are classified as fast responders in the presence of noise (Fig. S2).

Confident that the mathematical model can capture the observed range of responses, we proceed to train the model by sampling from the posterior for each of the 40 patients in the training set. The full posterior, formed by concatenating equal numbers of posterior samples from each patient in the training set (Eq. 7), is shown alongside

the prior in Fig. 3. Note that the full posterior represents parameter combinations that can be attributed to patients throughout the population (the parameters vary patient-to-patient), and does not represent uncertainty in each parameter within any individual patient. Therefore, we are less interested in whether such parameters are identifiable, but rather that the full posterior now contains knowledge about the set of patient responses observed in the training set.

The correlation structure in the joint posterior is extremely important: marginal densities provide little information about each parameter and produce meaningless predictions when sampled independently. Therefore, in Fig. 4 we investigate the correlation structure by examining the set of pair-wise bivariate marginal distributions. To gauge how parameter combinations vary with each radiotherapy response classification, we classify each posterior sample into a response class based on the criteria set out in Sect. 2.2.1. The proportion of samples attributed to each class is shown in Table 1.

First, it is evident from results in Fig. 4 that the predicted value of the initial necrotic proportion, ϕ_0, does not vary between fast and poor responders. This is seen in bivariate denisties between ϕ_0 and all other parameters. The statistic does, however, appear to distinguish pseudo-progressors from the other response types: estimates for ϕ_0 suggest that tumours in such patients contain a much larger necrotic region pre-treatment. Faster responders are characterised in relation to poor responders by both a higher radiotherapy sensitivity, γ, and necrotic material decay rate, ζ. The necrotic material decay rate also appears to distinguish poor, fast, and plateaued responders: poor responders through a very low decay rate, plateaued responders by a high decay rate, and fast responders an intermediate rate. Finally, results in Fig. 4 suggest that pseudo-progressors are characterised by both a high cell proliferation rate and correspondingly high radiotherapy response.

3.2 Model Predictions

Given that the training set is relatively small, a potential obstacle is that responses of new patients may not be similar enough to those of existing patients to produce reliable predictions; indeed only 6.0% of posterior samples correspond to patients that exhibit a poor response to treatment. To address this with the existing data, we "expand" the full posterior to form the second-level prior, $p_2(\boldsymbol{\theta})$, by resampling and perturbing (essentially, forming $p_2(\boldsymbol{\theta})$ as a multivariate kernel density estimate based on the full posterior, with a kernel variance expanded from Silverman's rule to account for new patient dissimilarity). The updated proportions, based on 100,000 samples from $p_2(\boldsymbol{\theta})$, are given in Table 1 and suggest an updated prior probability of a new patient exhibiting a response at 65.3%. An alternative approach that is beyond the scope of the current work would be to stratify perturbed full posterior samples based on an external and accepted classification ratio: for example, to choose the prior weights $\{w_i\}$ to achieve a desired prior ratio of patients in each classification. These results highlight the difficulty of classifying patient outcomes based on a relatively small cohort of patients with little prior parameter knowledge. Using the first-level prior
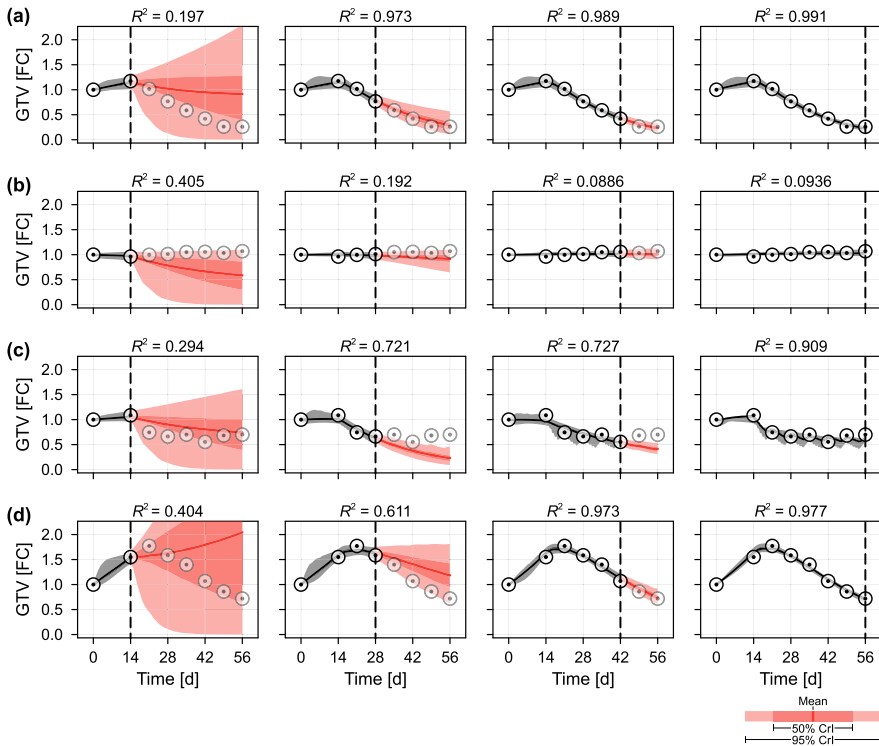
Fig. 5 Temporal predictions for four synthetic patients. Synthetic data from patients exhibiting **a** a fast response; **b** a poor response; **c** a plateaued response; and **d** pseudo-progression, are produced and used for predictions at various stages through the patient's treatment regime. In each case, the vertical dashed line indicates when the prediction is made: opaque marks indicate already-observed data used to produce predictions, semi-transparent marks indicate the future, as yet unobserved, trajectory. Predictions are represented as means (solid), 50% credible intervals (dark black or red shading), and 95% credible intervals (light black or red shading) constructed from weighted posterior samples. Model trajectories are coloured black (for retrospective predictions of tumour progression up to the present) and red (for prospective predictions of future tumour progression) (colour figure online)

(i.e., excluding all knowledge gained through analysis of the training data) further reduces the prior probability of an eventual response to 44.7%.

We first assess the predictive ability of our trained model by generating data from four synthetic patients exhibiting a fast response (Fig. 5a); a poor response (Fig. 5b); a plateaued response (Fig. 5c); and pseudo-progression (Fig. 5d). Given that each set of patient-specific parameters is resampled from the full posterior, we expect each synthetic patient to display a similar response to at least one patient in the training set. Additionally, as each set of synthetic data is generated by the mathematical model, we are guaranteed that the observed response is within the possible gamut of model responses. We provide a table summarising the parameter values used for each patient in the supplementary material (Table S1).

In Fig. 5 we simulate real-time predictions by calibrating and forming predictions each week throughout treatment (i.e., at the time of each weekly CT scan). We show
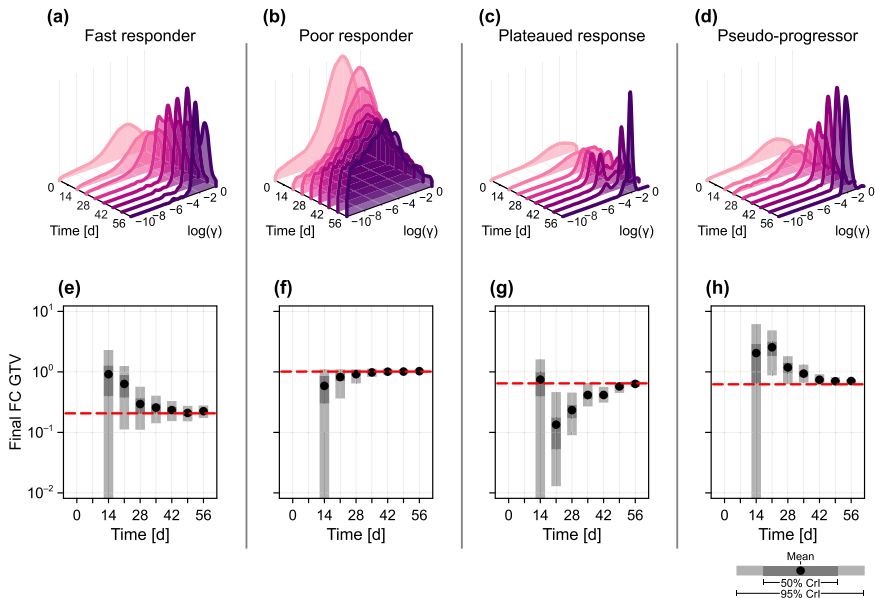
Fig. 6 Predictions for four synthetic patients. For the four patients analysed in Fig. 5 we show **a–d** the evolution of the posterior distribution relating to radiotherapy response, γ; and **e–h** the evolution of predictions for the relative tumour volume at the conclusion of treatment. In all cases, data up to, and including, the relevant time are included in the prediction. In **e–h**, we show the mean (black disc), and both 50% and 95% credible intervals for the final tumour volume, together with the true final tumour volume (red dashed), both given as the fold-change (FC) relative to the initial volume, $V(0)$. The true values of γ used for each patient are given as supplementary material (Table S1) (colour figure online)

predictions made at the start of treatment ($t = 14\,\mathrm{d}$), and every second week following ($t = 28\,\mathrm{d}$, $42\,\mathrm{d}$ and $56\,\mathrm{d}$). The results shown for $t = 56\,\mathrm{d}$ correspond to retrospective analysis of the trajectory, after all measurements have been taken, while the predictions drawn at $t = 14\,\mathrm{d}$ are made pre-treatment, before any radiotherapy response has been observed. As a class under-represented in the data set and hence the prior, predictions made for the pseudo-progressor at $t = 28\,\mathrm{d}$ almost entirely miss the true trajectory. Consequently, the single data point at $t = 28\,\mathrm{d}$ that sees a decrease is judged alongside both prior knowledge and potential measurement noise.

To quantitatively compare the time-evolution of prediction confidence, we plot in Fig. 6a–d the evolution of posterior information relating to the radiotherapy response, γ, and in Fig. 6e–h the time evolution of predicted final tumour volume (i.e., the fold-change GTV at $t = 56\,\mathrm{d}$ compared to the measurement at $t = 0\,\mathrm{d}$). The most immediate result is that both the fast responders and pseudo-progressors yield a posterior density for γ higher than that for the poor-responders. The results in Fig. 6e show that the predicted final GTV quickly narrows around the true value for the fast responder, but takes longer for the plateaued progressors and pseudo-responders. At the same time, the results in Fig. 6f show that by two weeks into treatment, the model predicts with 95% confidence that a patient will not see a final GTV less than 50% of that pre-treatment. The results in Fig. 6g highlight again the difficulties faced when
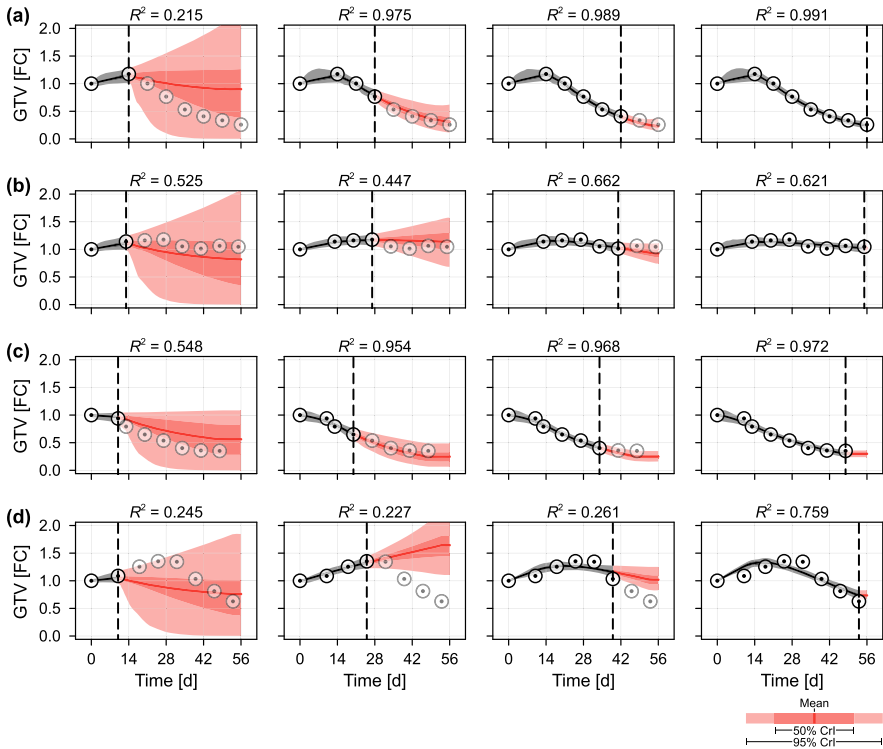
Fig. 7 Temporal predictions for the four patients excluded from the training set. We reproduce the analysis from Fig. 5 for the four patients in Fig. 1. These patients were not included in the training set, and so these results are representative of clinical predictions made throughout a new patient's course of treatment. Patients were classified previously as **a** a fast responder; **b** a poor responder; **c** exhibiting a plateaued response; and, **d** exhibiting pseudo-progression. Results related to the remaining seven patients excluded from the training set are given in the supplementary material (Fig. S6)

drawing predictions for patients exhibiting relatively rare responses: working with synthetic data eliminates the question of model-misspecification, however the 95% credible intervals produced from predictions drawn at $t = 21$ d and $t = 28$ d do not cover the true value (which can be calculated by resimulating data from each synthetic patient without measurement noise). Given GTV alone, it is not until $t = 42$ d (four weeks into treatment) that the model predicts with 95% confidence that the patient's tumour will eventually see a reduction in volume. This is in line with previous reports that mid-treatment responses correlate with outcome (Zahid et al. 2021b).

3.2.1 Clinical Data

Now that we have validated the model's ability to predict the time course of GTV for synthetic patients with a variety of radiotherapy responses, we turn to focus on drawing real-time predictions from unseen clinical data.

In Figs. 7 and 8, we repeat the analysis performed in Figs. 5 and 6 for the four patients initially exhibited in Fig. 1. We remind the reader that, although we previously
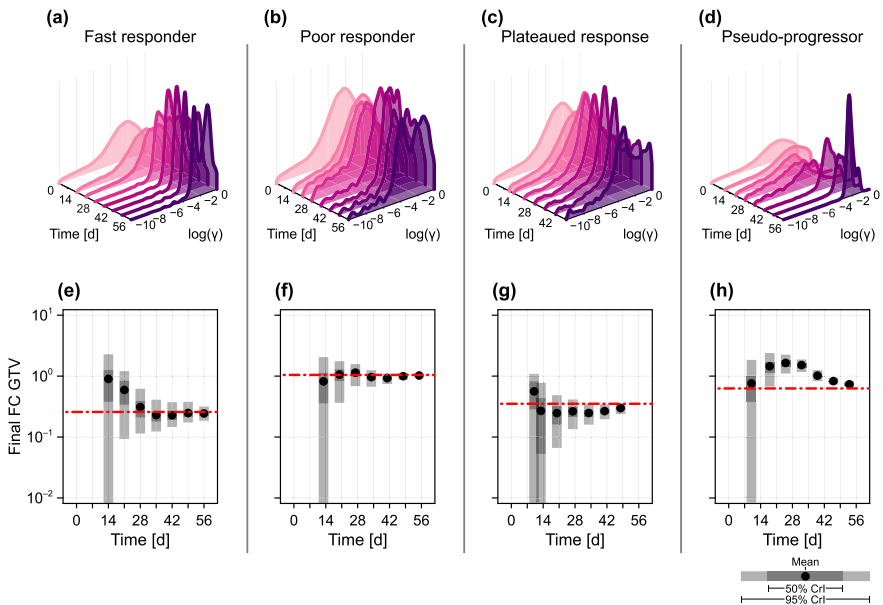
Fig. 8 Predictions for the four patients excluded from the training set. We reproduce the analysis from Fig. 6 for the four patients in Fig. 1b–e. These patients were not included in the training set, and so these results are representative of clinical predictions made throughout a new patient's course of treatment

demonstrated that the model can reproduce the clinical observations for these four patients, none were included in the training set. Hence, predictions drawn up to a particular time include only GTV data up to and including that time, and knowledge gained from the training set. For completeness, in the supplementary material we reproduce the results in Fig. 7 for all 51 patients using a leave-out-one-cross-validation approach, where predictions for each patient are drawn from a training set comprising the other 50 patients.

At the time of treatment onset ($t = 14$ d in Fig. 7a, b and $t = 12$ d in Fig. 7c, d), predicted trajectories are similar and predominantly represent prior knowledge from the training set. By day 28, for the fast responder, and day 21, for the patient that eventually exhibits a plateaued response, the model predicts with 95% confidence that the patient will eventually achieve an overall reduction in tumour volume. Indeed, for both of these patients the precision in predictions of the final tumour volume narrows quickly around what is eventually observed. In contrast, at day 28 the patient that eventually exhibits a poor response sees roughly half of all predicted trajectories indicating an eventual increase in volume, and half a decrease. Throughout treatment, the mean prediction remains around the eventually observed value of unity. The results for the pseudo-progressor mirror those observed in the synthetic data: the predictions are perhaps initially misleading due to the relatively small (2.3%) prior probability of a patient exhibiting such a response.

To quantitatively explore the model's ability to predict patient classification, in Fig. 9 we plot the posterior classification probabilities for predictions drawn at each time point, in addition to a pooled classification probability of a patient displaying

a response (i.e., not a poor responder). Initially, at $t = 0$ d, the classification probabilities represent those in the second-level prior, $p_2(\boldsymbol{\theta})$ (Table 1). The most notable results are for the relatively rare classifications of plateaued response and pseudo-progressor. In the case of the former, the patient has a posterior classification mode (i.e., the most likely classification given all the information collected during the patients' course of treatment) of a fast responder. This again highlights the difficulties distinguishing plateaued responses from observation noise seen in faster responders. The pseudo-progressor, however, begins to gain a correct posterior classification probability by $t = 42$ d, just over four weeks into treatment. The classifications following the first measurement at $t = 14$ d are qualitatively similar to that observed in the prior, subsequent measurements which show an increase in gross tumour volume lead to classification as a poor responder, highlighting the limitations of the currently trained model in distinguishing pseudo-progressors from poor responders.

To explore the relative value of existing and newly collected information, in the supplementary material we produce additional results that show temporal predictions for both synthetic and validation patients, produced using the uninformative (i.e., first-level) prior. These results correspond to a prior probability of 44.7% that a patient will eventually respond to treatment; much lower than that estimated from analysis of the training data (94.0%) and that in the second-level prior (65.3%). For the synthetic patients presented in Fig. 5, the results show a decrease in prediction fit (as measured by Bayesian R^2) for predictions drawn prior to $t = 28$ d. For times later than $t = 42$ d, predictions drawn using both the uninformative and informative priors are comparable. Similar results are seen for the validation patients presented in Fig. 7, although the differences are less pronounced from the third-post-radiotherapy observation point onwards. The difference between results for the synthetic and validation patients is expected: the informative prior is known to be representative for the synthetic patients, whereas we do not have this guarantee for the validation patients. Hence, observed information is more important than prior information in newly informed patients that are not well-represented by the prior.

3.3 Value in Collecting Measurements of Tumour Heterogeneity

The weekly GTV used for our analysis already exceeds clinical practice of just two pretreatment CT scans per patient. To assess the potential value of collecting higher-quality scan data that additionally enables identification of the tumour's necrotic volume, we repeat our analysis of the synthetic patient in Fig. 5a given that noisy measurements of both $V(t)$ and $N(t)$ are now available. The results in Fig. 10a, b show that, by day 28, relatively precise predictions relating to the trajectories of both variables can now be made. In Fig. 10c we quantitatively compare predictions for the final GTV in both scenarios. As expected, more precise estimates can be made should data relating to both variables be available.

In Fig. 11, we repeat the analysis for two new synthetic patients that experience a poor response. In the case of the first patient, a small gain in predictive ability is seen from the inclusion of necrotic volume measurements (Fig. 11c); interestingly, this improvement is not seen for the second patient (Fig. 11f). Overall, these results
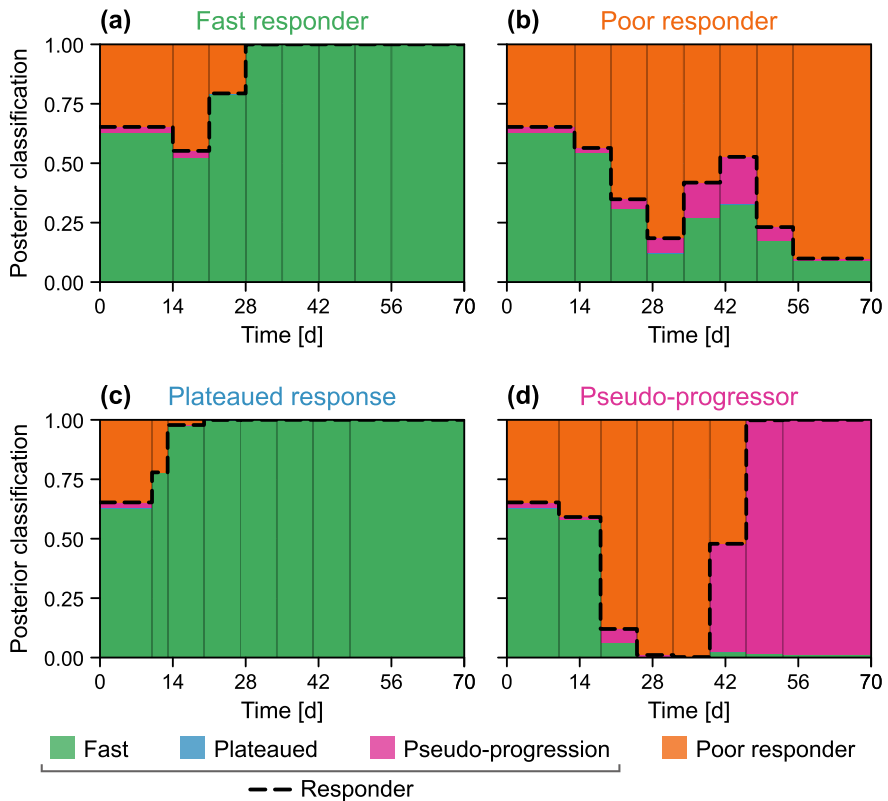
Fig. 9 Classification of the four patients excluded from the training set. We predict each patient's classified response using data up to and including the relevant time (height of each region indicates the predicted proportion). The predicted probability of the patient responding (i.e., receiving a classification that is not that of a poor responder) is shown in black dashed. Before the start of treatment, the predicted classifications correspond to those of the second-level prior in Table 2

highlight a key challenge with using the population-calibrated mathematical model to draw predictions relating to tumour composition and the underlying cause of a poor response, particularly given the wide-ranging spatial compositions seen in poor responders. The first synthetic patient exhibits a poor response due to the development of a tumour comprising almost entirely necrotic material, which does not degrade (Fig. 11b), while the tumour composition in the second synthetic patient is perhaps more realistic, with the necrotic fraction comprising approximately 60% of the GTV at the end of treatment. (Fig. 11e). Since the model is not trained using clinical data relating to tumour composition, it cannot distinguish between tumour compositions that are clinically realistic and those that are not. This is not an issue for prediction of the GTV, as prediction uncertainty incorporates all possible tumour compositions through prior knowledge. Predictions of necrotic volume, meanwhile, represent predominantly prior knowledge in addition to restrictions imposed by the modelled relationship between the observed GTV of patients in the training set and their potential inner tumour composition.
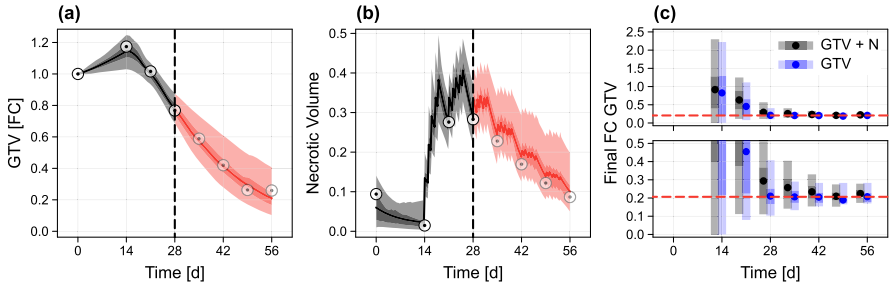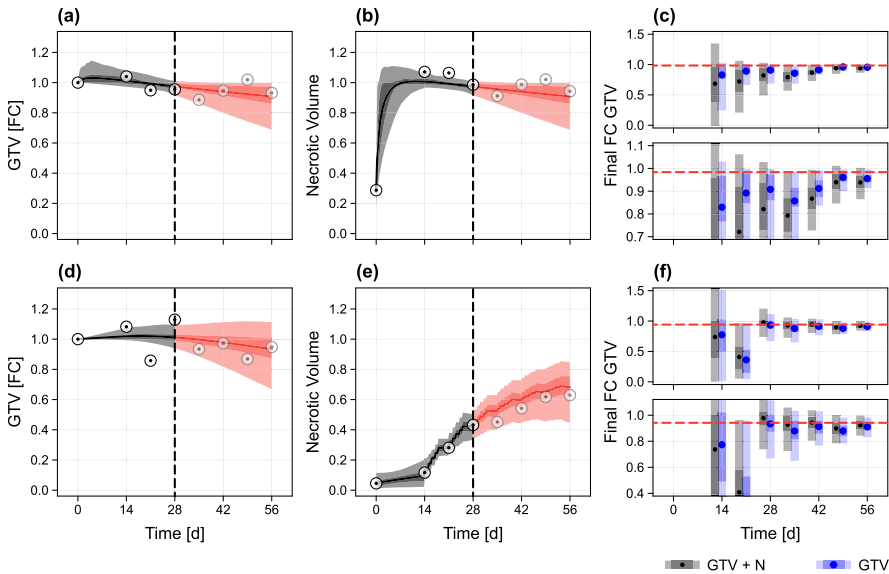
Fig. 10 Predictions for a synthetic patient with a fast response subject to both GTV and necrosis measurements. **a–b** We reproduce the analysis from Fig. 5a in the case that information relating to both $V(t)$ and $N(t)$ is available. **c** Mean, 50%, and 95% credible intervals for the final GTV in both data collection scenarios. The true value (calculated by resimulating data from each synthetic patient without measurement noise) is also shown (red dashed). Lower plot in **c** is a cropped inset of the upper (colour figure online)

Fig. 11 Predictions for two synthetic patients with a poor response subject to both GTV and necrosis measurements. **a–b**, **d–e** We produce dynamic predictions of tumour progression for each patient in the case that information relating to both $V(t)$ and $N(t)$ is available. **c**, **f** Mean, 50%, and 95% credible intervals for the final GTV in both data collection scenarios. The true value (calculated by resimulating data from each synthetic patient without measurement noise) is also shown (red dashed). Lower plot in each set is a cropped inset of the upper (colour figure online)

4 Conclusion

The development of predictive mathematical models of patient-specific tumour response is hindered by multiple challenges. Mathematical models must incorporate sufficient detail to capture a wide range of potential responses, while clinical data are highly limited, often comprising just one or two noisy measurements of tumour

volume prior to treatment initiation. Advances in imaging technologies or the use of magnetic resonance imaging embedded in radiation delivery devices may, in future, provide a cost-effective means of collecting more detailed information, allowing the calibration of correspondingly more detailed mathematical models (Gatenby et al. 2013; Gillies and Balagurunathan 2018; McGee et al. 2021; Park et al. 2023). In this work, however, we work with a fundamental set of measurements, and present the statistical methodology and an appropriately complex mathematical model to maximise data utility and draw clinically relevant predictions by leveraging a cohort of patients that exhibit a variety of treatment responses.

Importantly, the two compartment model is able to reproduce the full range of patient responses observed in our cohort of clinical data, representing an improvement over previously proposed one-compartment models which may not capture more complex behaviours, such as the plateaued response and pseudo-progressor behaviour. This is particularly important for prediction, since the choice of model and gamut of possible responses form a significant part of prior knowledge. While the mathematical literature presents an extensive catalogue of more complex models, we find that our choice of model with six unknown parameters, all with a direct biophysical interpretation, is simultaneously both sufficiently simple to ensure practical identifiability in some cases, and sufficiently complex to produce the variety of responses seen in the clinical data. Parameter identifiability is clearly not essential to produce predictions (single patient predictions drawn early in the course of treatment from the first-level prior, where the number of parameters exceeds the number of data points, are still sensical), however the relatively small parameter space and resultant tightly constrained second-level prior (Fig. 4) ensures adequate coverage in our resampling-based inference method: we expect our approach to become prohibitively expensive for models with large numbers of parameters.

The overarching goal of the presented framework is to leverage existing clinical data to produce a predictive model for GTV that accurately captures the uncertainty in predictions made for new patients. By benchmarking against both synthetic and a validation clinical data set, we show that our approach excels at this goal for patients with more typical responses: the fast and poor responders. Given the relatively small size of our training data—comprising measurements from 40 patients—it is no surprise that our approach does not perform as well for patients with atypical responses: pseudo-progressors, for instance, make up only 2.3% of the prior, meaning that the GTV progression of these patients is informed by (on average) a single patient in the training set. In this case, it takes six on-treatment measurements before the patient is identified as more likely to exhibit an eventual response than a poor response. The most effective remedy would be to accumulate significantly more clinical data with better representation of outliers. Should enough data become available, stratification could be used to ensure that representation of patients in the training data either concords with that in the population, or incorporates non-quantitative prior knowledge (such as patient characteristics) that pre-inform similarities with patients in the training set. Our modelling framework is well-poised to incorporate more detailed clinical data, including, for instance, radiotherapy plan adaptation and information relating to variations in delivered dose throughout the course of treatment. Inclusion of such information is

likely to lead to better response classification, particularly if the radiotherapy dose is modified during the course of treatment.

Both the accuracy and precision of predictions could also be improved for all patients through a better biological understanding of radiotherapy response. The final set of results presented in this work highlight that GTV measurements alone are insufficient to identify the root cause of a poor response. Indeed, predictions related to the inner tumour composition must be treated with as much caution as with predictions for atypical patients that are dissimilar to all patients in the training set. The absence of tumour composition data in the training set means that all predictions of tumour composition are only informed by data indirectly through the model, which has, in turn, been validated against solely GTV data. The prospect of training a model with joint GTV-composition measurements is at present hypothetical, although entirely possible through advanced imaging technologies (Sun et al. 2018; Salem et al. 2019; Rockne et al. 2019). At this stage, our framework could additionally be applied to answer important questions relating to the number of tumour composition measurements required to accurately predict patient outcome throughout their course of treatment.

We highlight that our statistical methodology is, for the most part, model agnostic. Thus, informed by more detailed data, our approach could be used to develop a fully validated predictive model of not just GTV, but tumour composition, cell density, proliferation, hypoxia, and more. However, this proposition is not without limitation: our current choice to bootstrap parameter samples is likely to perform poorly for models with a large number of parameters. Such dimensionality-induced issues can be in part alleviated by sampling the full posterior directly, although this would introduce additional computational challenges. Further statistical developments are also needed to include parameters that are fixed between patients (for example, the noise parameters), or parameters that are assumed to be uncorrelated to others.

Our results add to a growing body of work (Claret et al. 2009; Ribba et al. 2012; Rockne et al. 2019; Bruno et al. 2020) that highlights the utility that mathematical models could bring to the clinic; in future informed by highly detailed and representative patient data to provide objective, real-time, and personalised patient predictions that inform clinical decision-making.

Author Contributions All authors conceived the study, provided feedback on drafts, and gave approval for final publication. A.P.B. and T.L. drafted the manuscript. A.P.B. implemented the computational algorithms. J.C.ds collected and provided the clinical data.

Data availability Code used to produce the results are available on GitHub at https://github.com/ap-browning/clinical_predictions.

References

Alfonso JCL, Jagiella N, Núñez L et al (2014) Estimating dose painting effects in radiotherapy: a mathematical model. PLoS ONE 9(2):e89380. https://doi.org/10.1371/journal.pone.0089380

Araujo RP, McElwain DLS (2003) A history of the study of solid tumour growth: the contribution of mathematical modelling. Bullet Math Biol 66(5):1039. https://doi.org/10.1016/j.bulm.2003.11.002

Belgioia L, Morbelli SD, Corvò R (2021) Prediction of response in head and neck tumor: focus on main hot topics in research. Front Oncol 10:604965. https://doi.org/10.3389/fonc.2020.604965

Bobadilla AVP, Maini PK, Byrne H (2017) A stochastic model for tumour control probability that accounts for repair from sublethal damage. Math Med Biol J IMA 35(2):181–202. https://doi.org/10.1093/imammb/dqw024

Brady R, Enderling H (2019) Mathematical models of cancer: when to predict novel therapies, and when not to. Bullet Math Biol 81(10):3722–3731. https://doi.org/10.1007/s11538-019-00640-x

Browning AP, Simpson MJ (2023) Geometric analysis enables biological insight from complex non-identifiable models using simple surrogates. PLoS Comput Biol 19(1):e1010844. https://doi.org/10.1371/journal.pcbi.1010844

Browning AP, Sharp JA, Murphy RJ et al (2021) Quantitative analysis of tumour spheroid structure. eLife 10:e73020. https://doi.org/10.7554/elife.73020

Bruno R, Bottino D, De Alwis DP et al (2020) Progress and opportunities to advance clinical cancer therapeutics using tumor dynamic models. Clin Cancer Res 26(8):1787–1795. https://doi.org/10.1158/1078-0432.ccr-19-0287

Caudell JJ, Torres-Roca JF, Gillies RJ et al (2017) The future of personalised radiotherapy for head and neck cancer. Lancet Oncol 18(5):266–273. https://doi.org/10.1016/s1470-2045(17)30252-8

Chvetsov AV (2013) Tumor response parameters for head and neck cancer derived from tumor-volume variation during radiation therapy. Med Phys 40(3):034101. https://doi.org/10.1118/1.4789632

Chvetsov AV, Dong L, Palta JR et al (2009) Tumor-volume simulation during radiotherapy for head-and-neck cancer using a four-level cell population model. Int J Radiat Oncol 75(2):595–602. https://doi.org/10.1016/j.ijrobp.2009.04.007

Chvetsov AV, Yartsev S, Schwartz JL et al (2014) Assessment of interpatient heterogeneity in tumor radiosensitivity for nonsmall cell lung cancer using tumor-volume variation data. Med Phys 41(61):064101. https://doi.org/10.1118/1.4875686

Claret L, Girard P, Hoff PM et al (2009) Model-based prediction of phase III overall survival in colorectal cancer on the basis of phase II tumor dynamics. J Clin Oncol 27(25):4103–4108. https://doi.org/10.1200/jco.2008.21.0807

Collis J, Connor AJ, Paczkowski M et al (2017) Bayesian calibration, validation and uncertainty quantification for predictive modelling of tumour growth: a tutorial. Bullet Math Biol 79(4):939–974. https://doi.org/10.1007/s11538-017-0258-5

Enderling H, Park D, Hlatky L et al (2009) The importance of spatial distribution of stemness and proliferation state in determining tumor radioresponse. Math Model Natl Phenom 4(3):117–133. https://doi.org/10.1051/mmnp/20094305

Enderling H, Alfonso JCL, Moros E et al (2019) Integrating mathematical modeling into the roadmap for personalized adaptive radiation therapy. Trends Cancer 5(8):467–474. https://doi.org/10.1016/j.trecan.2019.06.006

Fowler JF (2006) Development of radiobiology for oncology–a personal view. Phys Med Biol 51(13):263–286. https://doi.org/10.1088/0031-9155/51/13/r16

Gao X, McDonald JT, Hlatky L et al (2013) Acute and fractionated irradiation differentially modulate glioma stem cell division kinetics. Cancer Res 73(5):1481–1490. https://doi.org/10.1158/0008-5472.can-12-3429

Gatenby RA, Grove O, Gillies RJ (2013) Quantitative imaging in cancer evolution and ecology. Radiology 269(1):8–14. https://doi.org/10.1148/radiol.13122697

Gelman A, Goodrich B, Gabry J et al (2019) R-squared for Bayesian regression models. Am Stat 73(3):307–309. https://doi.org/10.1080/00031305.2018.1549100

Gillies RJ, Balagurunathan Y (2018) Perfusion MR imaging of breast cancer: insights using "habitat imaging". Radiology 288(1):36–37. https://doi.org/10.1148/radiol.2018180271

Gong J, Santos MMD, Finlay C et al (2011) Are more complicated tumour control probability models better? Math Med Biol 30(1):1–19. https://doi.org/10.1093/imammb/dqr023

Greenspan HP (1972) Models for the growth of a solid tumor by diffusion. Stud Appl Math 51(4):317–340. https://doi.org/10.1002/sapm1972514317

Hanin LG (2004) A stochastic model of tumor response to fractionated radiation: limit theorems and rate of convergence. Math Biosci 191(1):1–17. https://doi.org/10.1016/j.mbs.2004.04.003

Harshe I, Enderling H, Brady-Nicholls R (2023) Predicting patient-specific tumor dynamics: how many measurements are necessary? Cancers 15(5):1368. https://doi.org/10.3390/cancers15051368

Kreutz C, Raue A, Timmer J (2012) Likelihood based observability analysis and confidence intervals for predictions of dynamic models. BMC Syst Biol 6(1):120. https://doi.org/10.1186/1752-0509-6-120

Lawson BAJ, Drovandi CC, Cusimano N et al (2018) Unlocking data sets by calibrating populations of models to data density: a study in atrial electrophysiology. Sci Adv 4(1):e1701676. https://doi.org/10.1126/sciadv.1701676

Lewin T, Kim J, Latifi K et al (2016) Proliferation saturation index predicts oropharyngeal squamous cell cancer gross tumor volume reduction to prospectively identify patients for adaptive radiation therapy. Int J Radiat Oncol Biol Phys 94(4):903. https://doi.org/10.1016/j.ijrobp.2015.12.116

Lewin TD, Maini PK, Moros EG et al (2018) The evolution of tumour composition during fractionated radiotherapy: implications for outcome. Bullet Math Biol 80(5):1207–1235. https://doi.org/10.1007/s11538-018-0391-9

Lewin TD, Byrne HM, Maini PK et al (2020) The importance of dead material within a tumour on the dynamics in response to radiotherapy. Phys Med Biol 65(1):015007. https://doi.org/10.1088/1361-6560/ab4c27

McAneney H, O'Rourke SFC (2007) Investigation of various growth mechanisms of solid tumour growth within the linear-quadratic model for radiotherapy. Phys Med Biol 52(4):1039–1054. https://doi.org/10.1088/0031-9155/52/4/012

McGee KP, Hwang K, Sullivan DC et al (2021) Magnetic resonance biomarkers in radiation oncology: the report of AAPM Task Group 294. Med Phys 48(7):e697–e732. https://doi.org/10.1002/mp.14884

Moral PD, Doucet A, Jasra A (2006) Sequential Monte Carlo samplers. J R Stat Soc Ser B Stat Methodol 68(3):411–436. https://doi.org/10.1111/j.1467-9868.2006.00553.x

Park JC, Song B, Liang X et al (2023) A high-resolution cone beam computed tomography (HRCBCT) reconstruction framework for CBCT-guided online adaptive therapy. Med Phys. https://doi.org/10.1002/mp.16734

Poleszczuk J, Walker R, Moros EG et al (2018) Predicting patient-specific radiotherapy protocols based on mathematical model choice for proliferation saturation index. Bull Math Biol 80(5):1195–1206. https://doi.org/10.1007/s11538-017-0279-0

Powathil GG, Adamson DJA, Chaplain MAJ (2013) Towards predicting the response of a solid tumour to chemotherapy and radiotherapy treatments: clinical insights from a computational model. PLoS Comput Biol 9(7):e1003120. https://doi.org/10.1371/journal.pcbi.1003120

Powathil GG, Munro AJ, Chaplain MA et al (2016) Bystander effects and their implications for clinical radiation therapy: insights from multiscale in silico experiments. J Theor Biol 401:1–14. https://doi.org/10.1016/j.jtbi.2016.04.010

Prokopiou S, Moros EG, Poleszczuk J et al (2015) A proliferation saturation index to predict radiation response and personalize radiotherapy fractionation. Radiat Oncol 10(1):159. https://doi.org/10.1186/s13014-015-0465-x

Ribba B, Colin T, Schnell S (2006) A multiscale mathematical model of cancer, and its use in analyzing irradiation therapies. Theor Biol Med Model 3(1):7. https://doi.org/10.1186/1742-4682-3-7

Ribba B, Kaloshi G, Peyre M et al (2012) A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. Clin Cancer Res 18(18):5071–5080. https://doi.org/10.1158/1078-0432.ccr-12-0084

Richard M, Kirkby K, Webb R et al (2007) A mathematical model of response of cells to radiation. Nucl Instr Methods Phys Res Sect B Beam Inter Mater Atoms 255(1):18–22. https://doi.org/10.1016/j.nimb.2006.11.077

Rockne RC, Frankel P (2017) Mathematical modeling in radiation oncology. In: Wong JYC, Schultheiss TE, Radany EH (eds) Advances in radiation oncology. Cancer Treatment and Research, pp 255—271. https://doi.org/10.1007/978-3-319-53235-6

Rockne R, Alvord EC, Rockhill JK et al (2009) A mathematical model for brain tumor response to radiation therapy. J Math Biol 58(4–5):561. https://doi.org/10.1007/s00285-008-0219-6

Rockne R, Rockhill JK, Mrugala M et al (2010) Predicting the efficacy of radiotherapy in individual glioblastoma patients in vivo: a mathematical modeling approach. Phys Med Biol 55(12):3271–3285. https://doi.org/10.1088/0031-9155/55/12/001

Rockne RC, Trister AD, Jacobs J et al (2015) A patient-specific computational model of hypoxia-modulated radiation resistance in glioblastoma using 18F-FMISO-PET. J R Soc Interf 12(103):20141174. https://doi.org/10.1098/rsif.2014.1174

Rockne RC, Hawkins-Daarud A, Swanson KR et al (2019) The 2019 mathematical oncology roadmap. Phys Biol 16(4):041005. https://doi.org/10.1088/1478-3975/ab1a09

Sachs R, Hlatky L, Hahnfeldt P (2001) Simple ODE models of tumor growth and anti-angiogenic or radiation treatment. Math Comput Model 33(12–13):1297–1305. https://doi.org/10.1016/s0895-7177(00)00316-2

Salem A, Little RA, Latif A et al (2019) Oxygen-enhanced MRI is feasible, repeatable, and detects radiotherapy-induced change in hypoxia in xenograft models and in patients with non-small cell lung cancer. Clin Cancer Res 25(13):3818–3829. https://doi.org/10.1158/1078-0432.ccr-18-3932

Scott JG, Berglund A, Schell MJ et al (2017) A genome-based model for adjusting radiotherapy dose (GARD): a retrospective, cohort-based study. Lancet Oncol 18(2):202–211. https://doi.org/10.1016/s1470-2045(16)30648-9

Sharma S, Bekelman J, Lin A et al (2016) Clinical impact of prolonged diagnosis to treatment interval (DTI) among patients with oropharyngeal squamous cell carcinoma. Oral Oncol 56:17–24. https://doi.org/10.1016/j.oraloncology.2016.02.010

Stevens C, Bondy SJ, Loblaw DA (2013) Wait times in prostate cancer diagnosis and radiation treatment. Canad Urol Assoc J 4(4):243–8. https://doi.org/10.5489/cuaj.873

Sun Y, Reynolds HM, Wraith D et al (2018) Voxel-wise prostate cell density prediction using multiparametric magnetic resonance imaging and machine learning. Acta Oncol 57(11):1540–1546. https://doi.org/10.1080/0284186x.2018.1468084

Sunassee ED, Tan D, Ji N et al (2019) Proliferation saturation index in an adaptive Bayesian approach to predict patient-specific radiotherapy responses. Int J Radiat Biol 95(10):1421–1426. https://doi.org/10.1080/09553002.2019.1589013

Tariq I, Chen T, Kirkby NF et al (2016) Modelling and Bayesian adaptive prediction of individual patients tumour volume change during radiotherapy. Phys Med Biol 61(5):2145–2161. https://doi.org/10.1088/0031-9155/61/5/2145

Torres-Roca JF (2012) A molecular assay of tumor radiosensitivity: a roadmap towards biology-based personalized radiation therapy. Personal Med 9(5):547–557. https://doi.org/10.2217/pme.12.55

Vihola M (2020) Ergonomic and reliable Bayesian inference with adaptive Markov chain Monte Carlo. Wiley StatsRef: Statistics Reference Online pp 1–12. https://doi.org/10.1002/9781118445112.stat08286

Wang P, Feng Y (2013) A mathematical model of tumor volume changes during radiotherapy. Sci World J 2013:181070. https://doi.org/10.1155/2013/181070

Wang L, Correa CR, Hayman JA et al (2009) Time to treatment in patients with stage III non-small cell lung cancer. Int J Radiat Oncol 74(3):790–795. https://doi.org/10.1016/j.ijrobp.2008.08.039

Welch BL (1947) The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34(1/2):28. https://doi.org/10.2307/2332510

Yankeelov TE, Atuegwu N, Hormuth D et al (2013) Clinically relevant modeling of tumor growth and treatment response. Sci Transl Med 5(187):1879. https://doi.org/10.1126/scitranslmed.3005686

Zahid MU, Mohamed ASR, Caudell JJ et al (2021) Dynamics-adapted radiotherapy dose (DARD) for head and neck cancer radiotherapy dose personalization. J Personal Med 11(11):1124. https://doi.org/10.3390/jpm11111124

Zahid MU, Mohsin N, Mohamed AS et al (2021) Forecasting individual patient response to radiation therapy in head and neck cancer with a dynamic carrying capacity model. Int J Radiat Oncol 111(3):693–704. https://doi.org/10.1016/j.ijrobp.2021.05.132

Zaider M, Minerbo GN (2000) Tumour control probability: a formulation applicable to any temporal protocol of dose delivery. Phys Med Biol 45(2):279–293. https://doi.org/10.1088/0031-9155/45/2/303