University of Nottingham
UK | CHINA | MALAYSIA

Characterisation of glioblastoma sub-populations using mathematical modelling and inference

Bethan Edith Morris
4312831

Thesis submitted to The University of Nottingham for the degree of Doctor of Philosophy

Supervised by Prof Markus Owen, Prof Matthew Hubbard, Dr Ruman Rahman, Prof Dorothee Auer and Mr Stuart Smith

School of Mathematical Sciences
University of Nottingham

Abstract

Glioblastomas (GBMs) are the most aggressive primary brain tumours and have no known cure. Each individual tumour comprises multiple sub-populations of genetically-distinct cells that may respond differently to targeted therapies and may contribute to disappointing clinical trial results. Image-localized biopsy techniques allow multiple biopsies to be taken during surgery and provide information that identifies regions where particular sub-populations occur within an individual GBM, thus providing insight into their regional genetic variability. These sub-populations may also interact with one another in some way; it is important to ascertain the nature of these interactions, as they may have implications for responses to targeted therapies. In this work, we combine genetic information from image-localised biopsies with a mechanistic model of interacting GBM sub-populations to characterise the nature of interactions between two commonly occurring GBM sub-populations, those with EGFR and PDGFRA genes amplified.

Firstly, we develop a mathematical model using a PDE-based formalism and explore the dynamics of our model under a variety of interaction types (Chapter 2). Following on from this, we study population levels found across image-localized biopsy data from an initial cohort of patients and compare this to model outputs under competitive, cooperative and neutral interaction assumptions (Chapter 3). We explore other factors affecting the observed simulated sub-populations, such as selection advantages and phylogenetic ordering of mutations, and conduct a sensitivity analysis, as these factors may also contribute to the levels of EGFR and PDGFRA amplified populations observed in biopsy data.

The patient dataset is then expanded to include image-localised biopsies from additional patients and we examine the intra- and inter-tumoural heterogeneity in EGFR and PDGFRA amplification observed in this data (Chapter 4). We then proceed to explore the inferability of the model parameters using synthetic datasets. Finally, we perform inference for the patient dataset, where we are able to gain some insights into the dynamics of and nature of interactions between these amplified sub-populations.

Acknowledgements

Firstly, I would like to thank my supervisors in the mathematical sciences department, Professor Markus Owen and Professor Matthew Hubbard. Their guidance, advice and encouragement over the last few years has been invaluable and it has been a pleasure to work with and learn from them both.

I would also like to thank my co-supervisors, Professor Dorothee Auer, Mr Stuart Smith and Dr Ruman Rahman, for sharing their expertise of the biological, imaging and clinical aspects of glioblastomas with me.

I also wish to extend my thanks to Dr Kristin Swanson, Dr Andrea Hawkins-Daarud and Dr Lee Curtin from the Mathematical NeuroOncology Lab at the Mayo Clinic in Phoenix, Arizona. Firstly, for providing the patient data for this work and for the many meetings in which their ideas, support and encouragement were invaluable. And secondly, for facilitating two visits to their lab, where they made me feel so welcome and like a part of the team.

I would like to thank my examiners, Professor Bindi Brook and Professor Philip Maini, for taking the time to read this thesis, providing me with feedback and for making the viva such an enjoyable experience.

Finally, I would like to thank my family and friends: my parents for always supporting me throughout my education and encouraging me to take every opportunity; the friends I made in the maths department at Nottingham for making the PhD (and many climbing trips) so fun; and, finally, my husband for everything.

List of publications

Published work by the author that is included in this thesis, forming the basis of Chapter 3:

- B. Morris, L. Curtin, A. Hawkins-Daarud et al., "Identifying the spatial and temporal dynamics of molecularly-distinct glioblastoma sub-populations", *Mathematical Biosciences and Engineering* , 17(5), 4905–4941 (2020). Ref [68]

Other work by the author which is not explicitly included in this thesis:

- J. T. Nardini, J. H. Lagergren, A. Hawkins-Daarud, L. Curtin, B. Morris et al., "Learning equations from biological data with limited time samples", *Bulletin of Mathematical Biology* , 82(9), 1-33 (2020).

Contents

Abstract **2**

Acknowledgements **3**

1 Introduction and Background **1**

1.1 Glioblastoma 2

1.1.1 Genetic heterogeneity in glioblastomas 4

1.1.2 EGFR and PDGFRA amplification in glioblastomas 7

1.2 Existing approaches to modelling GBMs 9

1.2.1 The PI Model 9

1.2.2 The PIRT and PIHNA Models 12

1.2.3 Other Patient-Specific GBM Modelling 13

1.2.4 Models of tumour heterogeneity and adaptive therapies 14

1.3 Thesis overview and structure 17

2 Modelling glioblastomas with multiple interacting sub-populations 19

2.1 A model of interacting sub-populations in GBMs 19

2.2 Modelling the invasion of GBM cell populations 22

2.2.1 Derivation of a Multiple Species Cross-Diffusion Model 26

2.3 Travelling waves in the model of interacting sub-populations . . . 31

2.3.1 Phase Plane Analysis 31

2.3.2 Travelling Wave Analysis 44

2.4 Summary 54

3 Preliminary patient data and *in silico* investigations **57**

3.1 Introduction 57

3.2 A preliminary dataset of image-localised biopsies 58

3.3 Introducing EGFR and PDGFRA amplified sub-populations . . . 60
3.4 Results . . . 62
3.4.1 Our model predicts that distinct competing sub-populations of tumour cells can coexist in the same tumour region . . . 62
3.4.2 Simulation results . . . 64
3.4.3 LHS-PRCC Sensitivity Analysis . . . 73
3.5 Discussion . . . 78

4 Parameter Inference to characterise EGFR and PDGFRA amplified glioblastoma sub-populations 81
4.1 Introduction . . . 81
4.2 Image-localised biopsies from a cohort of GBM patients . . . 82
4.3 Introducing stochasticity in a simulated population of GBMs . . . 88
4.3.1 Defining a simulated cohort of GBMs . . . 91
4.4 Parameter Inference . . . 93
4.4.1 ABC-SMC Algorithm . . . 93
4.4.2 Inferring the parameters of the gamma distributions for N_E and N_P . . . 96
4.4.3 Inferring all model parameters using the ABC-SMC algorithm 110
4.5 Parameter inference for the patient data . . . 118
4.6 Discussion . . . 125

5 Conclusion 132

Bibliography 135

Appendices 153

A Finite Difference Scheme for Interacting Species Model in 1D 153

B Validation of travelling wave speeds 155

C Additional results exploring the factors affecting amplification patterns observed in simulations 157

D Effect of non-monotonicity of introduction locations in the LHS-PRCC sensitivity analysis 163

E Determining the vector of tolerances for the ABC-SMC algorithm 167

Chapter 1

Introduction and Background

Glioblastomas (GBMs) are the most common primary brain tumour occurring in adults and are a particularly aggressive form of cancer [56, 57]. Despite aggressive treatment protocols, there is no known cure and survival times for patients diagnosed with these tumours are poor at less than 15 months on average [111]. Clinicians are faced with several challenges when treating patients with GBMs, a key one being the intra-tumoural heterogeneity exhibited by these tumours. GBMs are known to be comprised of multiple genetically distinct tumour cells harbouring a variety of genetic mutations, which is thought to contribute to the failure of therapies, possibly through the survival of therapy-resistant tumour cells or the cooperation of sub-populations to evade therapy [5, 12, 72, 123]. Two sub-populations of interest in GBMs are those with amplification in the EGFR and PDGFRA genes.

In this thesis, we aim to characterise these sub-populations and gain insight into whether they may be interacting with one another in a cooperative manner. We present information from image-localised biopsies that provides insight into the distribution and co-occurrence of EGFR and PDGFRA amplified sub-populations throughout GBM tumours in a cohort of patients. While this provides important genetic and spatial information, this information is static and so it is difficult to extract any dynamic information about these tumour cells and any interactions that may be occurring between them. Mathematical modelling could be a useful tool in this scenario, as models can be used to enhance current knowledge and provide insight into complex biological processes. Therefore, we propose a novel mathematical model of interacting sub-populations and conduct *in silico* investigations and inference using patient image-localised biopsy data, with the aim of characterising the dynamics and nature of interactions of EGFR

and PDGFRA amplified sub-populations in GBMs.

Before moving onto this work, however, we provide some more detailed background information in this chapter. In Section 1.1, we begin by presenting an overview of GBMs and current treatment approaches and challenges, before discussing the heterogeneous nature of these tumours and efforts to characterise this through novel biopsy sampling techniques and the emerging field of radiogenomics. We then consider EGFR and PDGFRA amplification in GBMs and discuss their importance and why they are of particular interest in this work. In Section 1.2, we review the literature surrounding existing approaches to modelling GBMs and, finally, we summarise the key objectives and structure of this thesis in Section 1.3.

1.1 Glioblastoma

Glioblastomas are a type of glioma, a tumour arising from the cells surrounding neurons in the brain, called glial cells [37]. They are the most common primary brain tumour occurring in adults and are classed by the World Health Organisation as a grade IV glioma [56, 57], the most aggressive type, with a median survival time of a mere 4 months if left untreated [70]. Since its publication in 2005 [113], the Stupp Protocol has become the standard of care treatment for GBMs, which consists of maximal safe resection of the tumour, followed by radiation and the chemotherapy temozolomide. Despite this aggressive treatment, recurrent disease is inevitable and median survival times remain at just 14.6 months [111].

One of the primary reasons this treatment protocol ultimately fails is due to the diffuse nature of the disease; tumour cells infiltrate extensively throughout healthy brain tissue, far beyond the edge of the bulk tumour observed by magnetic resonance imaging (MRI) [37]. Tumour cells have been cultured from apparently healthy brain tissue as far as 4cm from the location of the bulk tumour [98]; Fig. 1.1 shows a schematic representation of tumour cells infiltrating far from the main bulk of the GBM. As a result of this, the region infiltrated by tumour cells is much larger than possible to remove during surgery or target with radiotherapy and, consequently, recurrence is inevitable as tumour cells are always left behind to proliferate and repopulate; recurrence has even occurred after hemispherectomies, the drastic removal of half of the brain [24, 31].

MRI is considered the "gold standard" for imaging brain tumours [92], however the true extent of infiltrative invasion of glioblastoma is often underestimated [17]. Figure 1.2 shows MRI scans for a patient showing an initial "complete" resec-

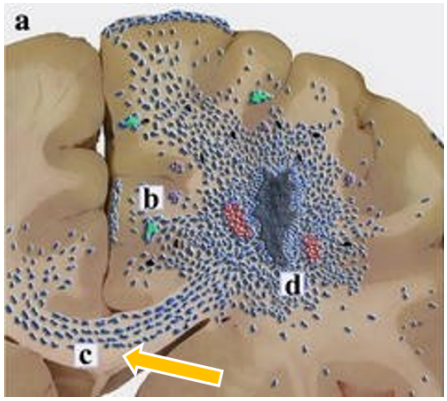


Figure 1.1: (a) Schematic diagram showing the infiltrative nature of GBMs; tumour cells are shown in blue, neurons in green (b) and vessels in red. (c) Tumour cells migrate across the corpus callosum (yellow arrow) to the other hemisphere of the brain. (d) The dark grey area shows a region of necrosis surrounded by tumour cells. Image from [17].

tion of the bulk tumour, followed by recurrence six months later and a subsequent second surgical removal, with the tumour recurring again three months later [37]. In spite of this inevitable failure of surgical resection, it is still often used as a treatment for GBM since it generally increases survival and quality of life [37], while more novel therapies for treating this disease are being investigated.

One of the novel approaches to treating patients with glioblastoma is targeted therapy, treatment that targets specific genes and proteins involved in the growth and survival of cancer cells. One such example is anti-angiogenic therapies targeting vascular endothelial growth factor (VEGF) [39]. VEGF is a protein that stimulates the formation of new blood vessels and has been found at high levels in GBM tumours [39, 43]; its overproduction is thought to partly explain damage to the blood brain barrier, oedema and regions of haemorrhage in GBMs [9, 39]. Therapies targeting VEGF have been tested extensively during various clinical trials, but have unfortunately not shown an improvement in overall survival (OS) of glioblastoma patients [39]. Other therapies have targeted various intracellular signalling pathways, inhibition of growth factor receptors [90, 126] and inhibition of integrins [34, 91, 112], all of which showed minimal or no efficacy [39]. Failure of these therapies could be due to the blood brain barrier preventing these agents targeting their pathways in GBM effectively if they are unable to reach their destination in sufficient concentrations [39].

While the ability of drugs to cross the blood brain barrier is a known challenge with treating GBMs—many research efforts are working to overcome this by developing novel techniques for delivery to the brain (see reference [106], for example)—some drugs, such as temozolomide, do still prove to be beneficial in

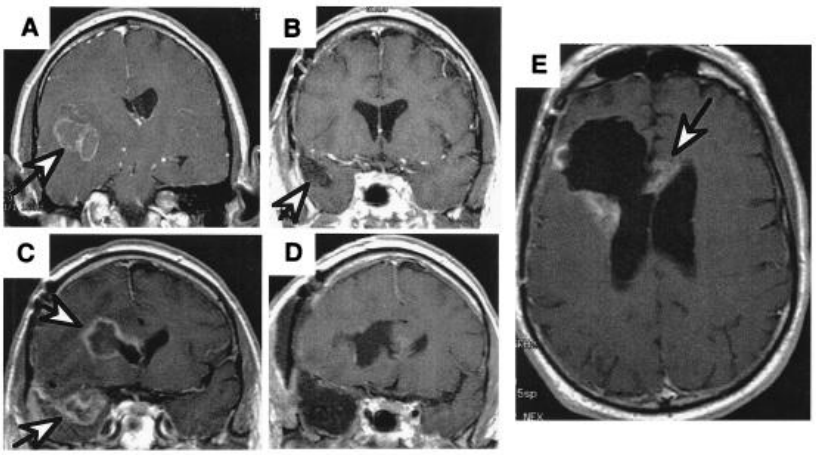


Figure 1.2: MRI scans of a patient showing recurrence after treatment. Scans taken in the coronal plane: (A) Pre-treatment scan showing GBM (arrow), (B) post-surgery scan showing clear resection cavity (arrow), (C) scan six months after surgery shows two recurrence sites (two arrows) and (D) scan shows resection cavities after second surgery. Final scan (E), taken three months later, shows tumour recurrence at the edge of the resection cavity and across the corpus callosum to the other hemisphere (arrow). Image from [37].

the treatment of these tumours [111]. Another major challenge in treating GBM is its genetic heterogeneity, which is thought to be an underlying cause for the failure of many targeted therapies [85, 108].

1.1.1 Genetic heterogeneity in glioblastomas

Glioblastoma is a disease characterised by inter- and intra-tumoural heterogeneity; each tumour is known to be comprised of several genotypically and phenotypically distinct sub-populations of tumour cells, with the sub-populations present varying both between patients and regions of the same tumour [12]. This heterogeneity is thought to contribute to the range of responses to therapies observed as different sub-populations may respond to a given therapy differently, with some tumours responding well, others to a lesser extent or partially and some not at all [103]. For example, sub-population ratios in tumour spheroids comprising multiple genetically distinct GBM cell lines were shown to change following treatment with various drugs, indicating that particular sub-populations were sensitive to some therapies, while others were not [102]. Further to this, the intra-tumoural heterogeneity of glioblastoma is thought to result in the failure of some therapies, as resistant sub-populations may be present within the tumour or different populations of tumour cells may cooperate to evade the therapy [5, 12, 72, 123]. As these scenarios would ultimately give rise to the therapy failing, it is important to gain a deeper understanding of the the sub-populations present in GBMs and

their implications for therapy, with the hope of being able to identify new targets for therapy and subgroups of patients whose tumours are likely to elicit the best response to particular therapies.

Over recent years, advances in biopsy sampling techniques have been helping to characterize this inter- and intra-tumoural genetic heterogeneity exhibited by GBMs. Treatment decisions are typically based on the biomarkers present within a single biopsy specimen, which may only be representative of a small part of the tumour. This means that clinical decisions made based on this knowledge may not be the optimal therapeutic strategy to target the majority of cancerous cells composing the rest of the tumour, which may have different genetic features and respond better to another treatment option. Therefore, in order to gain more understanding of the genetic heterogeneity across the tumour region, image-localized multiple biopsy sampling techniques have been developed, allowing surgeons to collect multiple tissue samples from an individual during surgery and record information about the location in the tumour from where the sample was taken. Subsequent tissue analysis then identifies how the genetic profile of these samples differs from one region of the tumour to another, providing spatial and genetic information that gives insight into the inter- and intra-tumoural heterogeneity present in these complex tumours; examples of such techniques being employed can be found in [41, 105, 108].

A further area of active research with regards to glioblastoma characterisation is the field of radiogenomics, where correlations between MRI patterns and genetic and cellular features of GBMs are analysed [23]; for a more detailed description of radiogenomics, the reader is referred to reference [44]. Various studies have shown that MR images of glioblastomas are influenced by the underlying genetic and cellular features of the tumour [8, 23], a simplified example of this would be the association of a given MRI texture feature with the presence of a particular tumour cell sub-population. This knowledge has led to the development of machine-learning models which are trained using genetic information from biopsies and sets of corresponding MRI features in an attempt to predict genetic or cellular features of a tumour from a patient's MRI scans. For example, until recently, IDH (isocitrate dehydrogenase) mutation status could only be determined from a biopsy sample taken during surgery. However, Zhang et al. were able to use a machine learning algorithm to predict IDH genotype in GBMs from pre-treatment clinical MRI features [134]. The ability to predict this non-invasively has potential prognostic utility, since mutations in this gene have been associated with longer overall survival in GBM patients compared to those with

the wildtype version [7, 83, 134]. Meanwhile, a study found that certain measures of texture analysis on pre-treatment MRIs to assess tumour heterogeneity were predictors of survival in a group of patients diagnosed with GBM [67], further highlighting the potential clinical utility of this field.

Furthermore, various work is being undertaken to characterise the regional heterogeneity in GBMs using radiogenomic techniques. For example, using genetic data from multiple image-guided biopsies from 13 GBM tumours, Hu et al. [40] were able to identify imaging correlations for 6 common driver genes found in GBMs. Using this information, they were able to produce models to predict amplification, i.e. an increase in the number of copies, of some of the driver genes in different regions of a tumour with high accuracies [40]. Following on from this work, they were able to produce radiogenomic maps showing predicted regions of amplification in the EGFR gene [42], a common driver gene found to be amplified in GBMs; more detail about the importance of this gene will be given in the following section. An example of these maps are shown in Fig. 1.3 overlaid on two MRI slices, where the maps were found to correctly predict the amplification status of image-localised biopsies [42]. The radigenomic maps shown in this example also predict heterogeneity in EGFR amplification status across the tumour region. The ability to quantify which parts of the tumour have specific gene alterations could help to determine how well regions of the tumour respond to targeted therapies [40]; in this example, the radiogenomic maps could help to determine how well regions of the tumour amplified in EGFR respond to EGFR targeted therapies. It is hoped that further research in this area will be able to identify a broader range of genetic alterations in GBMs and their intra-tumoural heterogeneity, thus helping to understand more about these complex tumours.

While the exact mechanism by which such genetic heterogeneity arises in GBMs is currently unknown, several possible theories have been proposed to explain this. These include the theory of clonal evolution, where tumours are thought to evolve through a process of acquiring mutations and natural selection; the cancer stem cell (CSC) model, in which a small population of CSCs give rise to and maintain the tumour through self-renewal and producing phenotypically diverse daughter cells; or possibly some complementary combination of these two theories [12]. In addition to understanding how genetic heterogeneity arises in GBMs and other types of cancer, further understanding of how such heterogeneity is maintained is also needed. One suggested mechanism of heterogeneity maintenance is called interclonal cooperativity, where interactions between different tumour cell populations are thought to be important; the theory suggests that

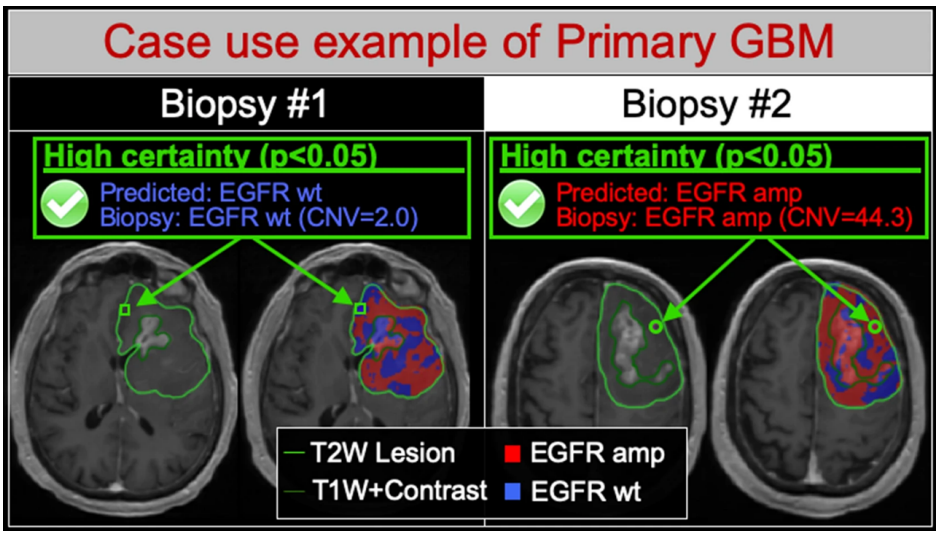


Figure 1.3: Two pairs of MRI slices are shown for an individual patient diagnosed with primary GBM. The location of two image-localised biopsies are shown, one on each of the slices, and the outline of the tumour burden observed through two different MRI modalities (T1 contrast enhanced and T2 weighted) are shown in dark and light green. On the right MRI of each pair, a radiogenomic map is overlaid showing predicted regions where the EGFR gene is amplified (red) and not amplified (blue). The maps were found to correctly predict the amplification status of each of the biopsies with a high certainty. Image reproduced from [42].

some cells may acquire mutations that result in the promotion of other tumour cell sub-populations in some way [12, 58]. One consequence of this could be that a small population of genetically-distinct tumour cells plays an important role in tumour progression and, thus, targeting such a sub-population of cells may have additional negative effects on the rest of the tumour cell population. Therefore, identifying such genetically distinct tumour cell sub-populations and understanding their interplay with other cell populations may have important implications for the success of therapies.

1.1.2 EGFR and PDGFRA amplification in glioblastomas

Two such populations of interest are those with amplification of the Epidermal Growth Factor Receptor (EGFR) and the Platelet-Derived Growth Factor Receptor Alpha (PDGFRA) genes, i.e., cells with an increased number of copies of the genes encoding each protein. While amplification status relates to the number of copies of a particular gene that a cell has in its DNA, its *copy number aberration*, it also induces overexpression of these genes in tumour cells [55]. The EGFR and PDGFRA proteins are both members of the Receptor Tyrosine Kinase (RTK) family of cell surface receptors which bind to a variety of growth factors,

cytokines and hormones and play a crucial role in the regulation of the signalling that controls cell proliferation, metabolism and survival [30]. Specifically, EGFR is a receptor that, upon binding, results in the activation of pathways that lead to cell proliferation, DNA synthesis and the expression of certain oncogenes [129] and its amplification has been shown to promote invasion in GBMs [81, 124] and be an unfavourable predictor for patient survival [97]. Meanwhile, PDGFRA is a receptor that, when bound, activates signalling pathways that promote oncogenesis [1, 11]. Due to the prevalence of EGFR and PDGFRA amplified tumour cells in GBMs—occurring in 41% and 10% of GBM samples in The Cancer Genome Atlas (TCGA) database, respectively [125]—these sub-populations have become prime molecular targets for therapies and a number of inhibitor drugs have been developed for this purpose [72]. These therapies targeting EGFR and PDGFRA amplified cells, however, have had limited success in GBMs in clinical trials so far [72].

Several possible mechanisms of chemoresistance to these drugs in GBMs are discussed by Nakada et al. [72]. However, one possible mechanism of chemoresistance to EGFR and PDGFRA targeted therapies of interest is through the interaction of cell sub-populations with amplification of these genes; these cells may interact in a cooperative way that facilitates their survival or, conversely, competitively, such that the targeting of one population with therapy benefits the other by removing its competitor. While the interactions between EGFR and PDGFRA amplified sub-populations are currently not well understood, it has been suggested that these cell populations may be interacting in a cooperative manner. For example, in experiments by Szerlip et al. [123] a form of cooperativity was observed between these cell populations, as combined inhibition of both receptors was needed to block activity of the PI3 kinase pathway—a pathway involved in the regulation of cell proliferation, apoptosis and migration [53]—in a mixed population of EGFR and PDGFRA amplified cells *in vitro*. In addition to this, Snuderl et al. [107] observed coexistence of these amplified sub-populations and suggested that they may co-evolve with similar fitness levels rather than compete during tumour evolution; the authors further suggest the possibility that these sub-populations cooperate to achieve a higher fitness level than each of the sub-populations individually [16, 107]. Little et al. [54] also observed similar co-existence patterns of distinct tumour cell sub-populations amplified in the EGFR and PDGFRA genes in GBM specimens, with each gene predominating in different areas of the same tumour specimen.

These observations about EGFR and PDGFRA amplified sub-populations

lead to questions about how such co-existence arises and what the possible implications for disease progression and treatment responses are. While suggestions of cooperative behaviour between EGFR and PDGFRA amplified tumour cells in GBMs have been proposed, the nature of any interactions between these sub-populations remains poorly understood and requires further study; in this thesis, we aim to use mathematical modelling approaches to gain an insight into possible interactions occurring between EGFR and PDGFRA amplified sub-populations in GBMs.

1.2 Existing approaches to modelling GBMs

It has been identified that new approaches to bring together existing clinical and scientific knowledge of GBMs are desperately needed as prognoses for patients with GBMs remain especially poor, despite the wealth of research in this area. Mathematical oncology has a role to play here, as models can be used to enhance current technologies and provide deeper insight. The term "mathematical oncology" here broadly means mathematically describing cancer using *in silico* models; these are models that use computers to simulate results and can reproduce the behaviour of a system using information obtained from clinical and experimental data [89]. Although a wide variety of models have been proposed to model glioblastoma [3], they generally fall into 2 categories: some describe tumour growth on the macroscopic scale, while others focus on tumour growth at the cellular level [89]. A recent comprehensive review of mathematical approaches to modelling GBMs is given by Alfonso et al. in [3], while Protopappa et al. provide a critical review from a clinician's perspective in [89] and an overview of mathematical modelling of cancer in general is given in [4]. Here we highlight some key approaches to modelling glioblastomas.

1.2.1 The PI Model

Over recent years, numerous papers on GBM modelling have been published by Kristin Swanson and her group at the Mayo Clinic in the USA. They have proposed a number of models of GBMs, the first—and most well-known—being the so called Proliferation-Invasion (PI) model [115, 116, 117, 118, 120, 121]. This model takes the form of a single equation, which is given by

$$\frac{\partial u}{\partial t} = \nabla \cdot (D \nabla u) + \rho u \left(1 - \frac{u}{K}\right), \tag{1.1}$$

where $u = u(\mathbf{x}, t)$ denotes the concentration of tumour cells, K denotes the maximum concentration and D and ρ are the diffusion coefficient and proliferation rate of the cells, respectively.

The model takes the form of the well known Fisher-KPP equation [28, 51] based on the simplified definition of cancer as "uncontrolled proliferation of cells with the ability to invade" [115] and uses the hypothesis that the "two final common pathways", proliferation and invasion, govern glioblastoma growth and capture the effects of any genetic-metabolic abnormalities occurring further upstream [115]. Although many more complex models of GBM growth exist, the real power of the PI model lies in its simplicity; it only has two model parameters which can be estimated from individual patient pre-treatment MRI scans [116, 118], enabling each patient with a glioblastoma to have an equation with their own parameter set that captures the growth kinetics of their particular tumour.

These model parameters, ρ and D, are estimated using three pre-treatment MRI scans—a T1Gd and a T2 weighted MRI taken at the same time, plus an additional scan of either type taken at a later time point. The "gradient" between this first pair of MRI scans can then be related to the ratio ρ/D [115] using observations that the circumference of tumour observed on a T1Gd weighted image represents the edge of solid tumour and the circumference observed on a T2 weighted MRI indicates a region of malignant cells existing at a lower concentration [49, 50]; these are hypothesized to correspond to tumour cell concentrations of 80% and 16% of the maximum concentration (K), respectively [115]. The third pre-treatment MRI scan taken at a later time point is then used to calculate the velocity of tumour radius expansion and Fisher's approximation, namely that the speed of front propagation tends asymptotically to $2\sqrt{\rho D}$, is utilized to obtain individual values of the parameters.

The model parameters, ρ and D, have been shown to be significantly associated with prognosis [130] and, using the PI model, Swanson et al. were able to make accurate survival predictions for patients following a range of surgical interventions [116] and identify retrospectively whether a given patient's tumour was sensitive or resistant to radiotherapy [115]. Furthermore, Baldock et al. found that a patient-specific metric of invasiveness derived from the PI model, given by ρ/D, predicts the survival benefit of gross total resection for GBM patients [6] and also the mutation status of the IDH1 gene in patients' tumours [7]. More recently, it has been shown that this patient-specific metric, ρ/D, also predicts overall survival following upfront radiotherapy with concurrent temozolomide [63].

Another application of the PI model has led to the creation of the Days Gained metric [73, 74]. By using the PI model to produce patient-specific simulations of untreated tumour growth, sometimes termed an "untreated virtual control (UVC)" [73], the "Days Gained" value is found as the difference in time between the post-treatment MRI scan and the time at which the UVC is predicted to have the same radius [4, 74]. This effectively compares the tumour size after treatment to the expected size of the tumour as if it had been left untreated [73]; Fig. 1.4 illustrates how the Days Gained score relates to the UVC and actual tumour size in a patient after receiving radiation therapy [74]. Unlike other response metrics, such as the Response Evaluation Criteria in Solid Tumours (RECIST) [26] and the Response Assessment in Neuro-Oncology (RANO)[132], Days Gained takes into account the variability in patients' tumour growth rates before treatment by using the PI model to determine patient-specific tumour growth rates. This enables the Days Gained metric to provide a more personalised assessment of treatment response, which other response metrics fail to do [73]. For example, a patient with a fast-growing glioblastoma that responds well to a therapy may show a pre-treatment-to-post-treatment change in tumour size on MRI that is similar to a patient with a slow-growing GBM that responds less well [73].

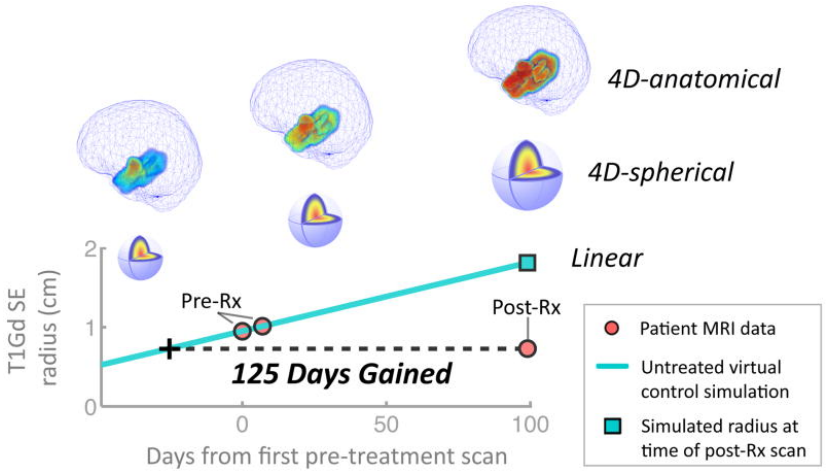


Figure 1.4: An illustration of the Days Gained score for a patient after receiving radiation therapy. Image reproduced from [74].

Neal et al. [73] also found that, using the Days Gained metric, they were able to distinguish true progression from pseudoprogression after radiation therapy; pseudoprogression is a "post-treatment radiation effect" where brain cells injured by the radiation cause a contrast-enhancing lesion to be visible on T1Gd MRI that looks like the tumour has recurred when it hasn't yet [84]. The Days Gained metric was also found to be prognostic for overall survival and progression free survival of glioblastoma patients [21, 74]. More recently, Days Gained has re-

portedly been shown to discriminate survival in patients receiving bevacizumab [21, 100], an anti-angiogenic drug, and gamma knife radiotherapy [21, 101], a treatment that enables clinicians to deliver intense radiation doses to a target area without damaging the surrounding tissue. Due to these results and the ease of use of the Days Gained metric, it is hoped that Days Gained scores will be used in a clinical setting in the future and help aid oncologists with clinical decision making [21].

1.2.2 The PIRT and PIHNA Models

Further to the PI model, Swanson's group have also published the PIRT (Proliferation-Invasion-Radiation-Therapy) and PIHNA (Proliferation-Invasion-Hypoxia-Necrosis-Angiogenesis) models of glioblastoma growth. The PIRT model incorporates the effects of radiotherapy into the patient-specific PI model and was used by Rockne et al. to predict radiation response in a cohort of patients with high accuracy [93]. Meanwhile, Corwin et al. used the PIRT model along with an multi-objective evolutionary algorithm to generate an optimized "patient-specific, biologically-guided" radiotherapy dose plan [19]. Upon comparison of this optimized dose plan with the standard of care, they demonstrated the potential to reduce the dose delivered to healthy brain tissue and produce a significant improvement in Days Gained scores for simulated tumour growths with treatment responses [19, 89].

The PIHNA model is an extension of the PI model that incorporates other micro-environmental factors, such as diffusible angiogenic factors, and multiple cell-type compartments, i.e. hypoxic, normoxic and necrotic cells. It was proposed with the aim of quantifying the role of angiogenesis in the progression of low-grade gliomas to high-grade, i.e. glioblastomas [122]. Swanson et al. found that the model described patterns of glioma growth dynamics visualised through MR imaging well and that the accumulation of genetic mutations in glioma cells were not necessary for progression from low- to high-grade [122]. A drawback of this more complex model, however, is the lack of patient-specificity due to the increased number of model parameters [122]. In spite of this, Gu et al. were able to produce patient-specific predictions of hypoxia throughout the tumour micro-environment by simulating FMISO-PET images, a type of imaging used to assess regions of low oxygen levels [35]; Fig. 1.5 shows a patient's FSIMO-PET image compared to the simulated version using the PIHNA model's prediction of hypoxia distribution [35].

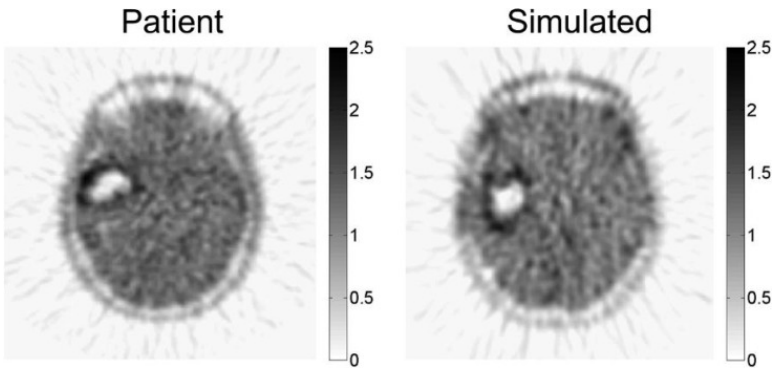


Figure 1.5: Patient FMISO-PET image compared with a simulated image using hypoxia regions predicted by the PIHNA model. Image reproduced from [35].

1.2.3 Other Patient-Specific GBM Modelling

Although all of the patient-specific glioblastoma modelling discussed so far has been published by Kristin Swanson and the Mathematical NeuroOncology Lab and collaborators, others are focussing their research on this area of modelling as well. A recent paper by Swan et al. [114] saw the application of the anisotropic diffusion model of Painter and Hillen [76] to a cohort of 10 glioma patients. This model is similar to the PI model, but, instead of considering isotropic diffusion (as in [116]), the authors derive patient specific anisotropic diffusion tensors informed by patient diffusion tensor images (DTIs) [114] to account for the known preferential migration of glioma cells along white matter tracts [86], the matter found deep in the brain that contains nerve fibres. By comparing Jaccard indices that measure the similarity between each of the patient-specific simulated tumour shapes (from the Painter and Hillen model and the PI model) and the shape of the actual tumour, it was found that incorporating anisotropic diffusion provided a slight improvement, but for patients with low anisotropy, the two models produced similar results as expected [114].

A model of glioma growth with anisotropic diffusion was also proposed by Jbabdi et al. [46], however only DTI data from a healthy individual was used so the diffusion tensor image derived didn't represent individual patients' brain architectures. Further to this, Patel and Hathout published a model that incorporated anisotropic diffusion and necrosis into a model of glioblastoma growth [86]; the diffusion tensor used was that derived by Jbabdi et al. [46] but scaled so the mean matches the patient-specific diffusion coefficient derived by Swanson et al. for the PI model in [116, 118]. The necrosis threshold and necrosis rate they use are defined from patient MRI scans based on the width of the enhancing rim of tumour surrounding the necrotic region [86]—a common feature of GBMs ob-

served on contrast enhanced T1 weighted MRIs—although it is not clear exactly how they assign these values. Using a mutual information metric, they conclude that their model produced improved simulated tumour progression profiles over previous models by Jbabdi et al. [46], which includes only anisotropic diffusion (with no necrosis), and the model by Woodward et al. [133] which includes neither anisotropic diffusion nor necrosis.

Another approach that utilises the width of the enhancing rim of tumour surrounding the necrotic region with the aim of personalising a mathematical model of GBM growth is by Pérez-Beteta et al. [88]. Their model consists of two coupled PDEs that reduce to the PI model [115] at tumour cell densities below a threshold and above which cell death is induced and necrosis forms [88]. They found that the enhancing rim width correlated negatively with survival, consistent with other studies. In their *in silico* investigations they find that the rim width correlates with tumour growth speed and hypothesize that the growth speed can be obtained by calculating the enhancing rim width from pre-treatment T1 weighted MRI scans of patients with GBM [88].

Hormuth et al. [38] also used multiple MRI scans to calibrate a family of reaction-diffusion models of high-grade glioma growth. In their work, they calibrated model parameters using imaging data for nine patients and used these to forecast spatially-mapped individual tumour response to chemo- and radiation-therapy at future imaging visits. They found that a novel two-species model describing the enhancing and non-enhancing tumour regions balanced model fit and complexity the best. This model was then used to predict future tumour growth and response at visits 3 and 5 months after radiotherapy, where they were able to predict the enhancing tumour volume with a low error at 3 months post-treatment.

1.2.4 Models of tumour heterogeneity and adaptive therapies

While models such as the PIHNA model [122] incorporate a form of tumour cell heterogeneity into the model formulation through considering populations of hypoxic, normoxic and necrotic cells, for example, they do not consider distinct sub-populations of cells with different phenotypes. Several modelling attempts have been made over recent years in order to capture the effects of intra-tumoural heterogeneity in glioblastoma on its growth and invasion dynamics and responses to therapy. For example, in order to explore the impact of clonal heterogeneity

on patterns of glioma growth, a 3D multi-scale agent-based model was proposed by Zhang et al. [135]. Incorporated in their model is a simple tumour progression pathway that leads to the emergence of 5 distinct tumour cell clones, each with different densities of EGFR (epidermal growth factor receptor) per cell. Additionally, they allow micro-environmental conditions to lead to different phenotypes of the clones, migratory, proliferative, apoptotic and quiescent, with the clone phenotype being determined by an EGFR gene-protein interaction [3]. Their results demonstrated that higher EGFR expression leads to faster expansion of the tumour region harbouring more aggressive cells due to a temporary competitive advantage, leading to asymmetric growth patterns similar to those observed clinically [3, 135]; Fig. 1.6 shows tumour cell clones migrating towards the nutrient source at different rates.

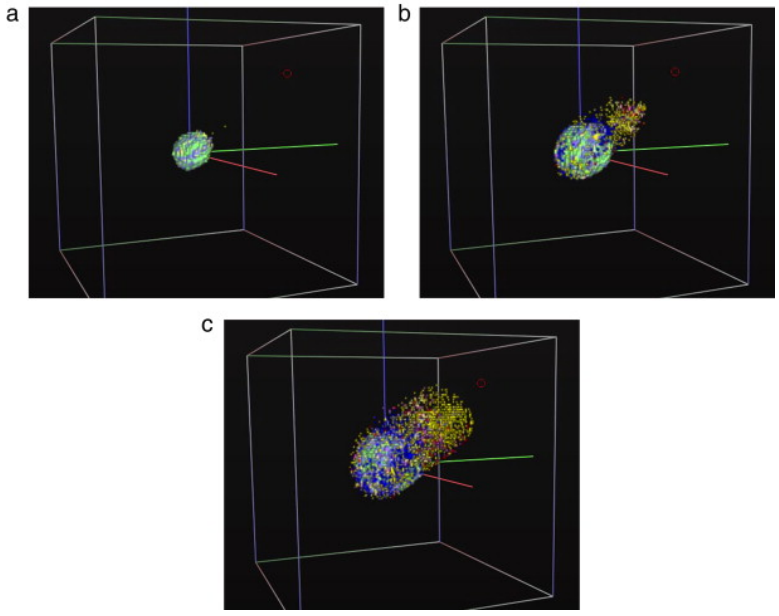


Figure 1.6: Snapshots of the simulated tumour growth, using the 3D multi-scale model proposed by Zhang et al. [135], at three different time points shows the tumour expanding and cells migrating towards the nutrient source (red circle). Each of the five colours green, blue, yellow, purple and red represent a different tumour cell clone. The light and dark grey colours represent cells in the quiescent and apoptosis states, respectively. Image reproduced from [135].

Frieboes et al. [29] proposed a 3D multi-scale model of tumour growth that incorporated several aspects of the tumour environment and two distinct tumour cell sub-populations; an original population and a mutated one. They explored the effects of different phenotypes of the mutated tumour cell population on the patterns of tumour growth. They found that an aggressive phenotype, characterised through increased substrate uptake and proliferation, led to the mutated cell population forming a ring surrounding the original tumour cell population

and a necrotic region. They also found that heterogeneity of oxygen, cell nutrients, and metabolites in the tumour microenvironment led to increased tumour instability leading to increased infiltration of surrounding healthy tissue. While not specifically modelling GBM tumours, this approach provides an interesting insight into possible links between heterogeneity in the tumour sub-populations and the tumour microenvironment with clinically observed tumour growth and invasion.

Whilst more understanding of intra-tumoural genetic heterogeneity in GBMs is needed, it is hoped that patterns of heterogeneity may be identified in the future and used to determine appropriate multi-modal therapeutic strategies on an individualised basis [108]. Several *in silico* mathematical modelling studies have explored what such a therapeutic strategy may look like. For example, Cunningham et al. [20] propose an experimental evolutionary therapy, which involves administering two therapies sequentially, with the idea being this will cause tumour cells to acquire specific adaptations leaving them vulnerable to the second therapy [20]. Another study used a model of two tumour cell populations, one drug-sensitive and the other resistant, to explore the effects of competition for space during adaptive therapy [110]. Using their model, Strobl et al. were able to visualise and quantify how treatment breaks during adaptive therapy increase the competitive inhibition of resistant cells. In particular, they examined how the spatial distribution of resistant and sensitive cells impacted the success of the adaptive therapy, where they found that it was most effective when resistant cells were clustered in a single location and surrounded by sensitive cells. They found that this was because inter-specific competition could be leveraged at the edge of the resistant colony to intra-specific competition between resistant cells at the core could be maximised [110]. Whilst these studies use *in silico* mathematical modelling, they serve as an example of how improved knowledge of intra-tumour heterogeneity has the potential to lead to improved GBM survival times through the development of novel approaches to treatment.

As mentioned previously, the approaches to modelling GBM mentioned here are just a few of the many works published and more comprehensive reviews are given in [3], [61] and [89]. However, it is clear from this brief overview that mathematical modelling has helped to provide some useful insights into the mechanisms underlying the complex disease that is glioblastoma. It is hoped that further modelling work will continue to bring together understanding behind clinical and experimental observations of GBMs and help to identify novel therapy targets [3], with the ultimate goal of finding an effective treatment for this disease.

1.3 Thesis overview and structure

As stated previously, if EGFR and PDGFRA amplified sub-populations are indeed interacting in GBMs, this will have implications for therapies targeting these cells and so it is important to gain more understanding of the nature of any interactions. The goal of this work, therefore, is to use mathematical modelling and inference to characterise the dynamics of these amplified sub-populations in GBMs and gain understanding of any interactions between them.

To this end, in Chapter 2, we present a novel mathematical model describing the growth of three distinct tumour cell sub-populations in GBMs; these are a population amplified in the EGFR gene, another amplified in the PDGFRA gene and a third sub-population amplified in neither gene. In our model, the two gene-amplified sub-populations interact with one another and we discuss the various different types of interactions that can occur. We then present a derivation of the terms in our model that describe the ability of each cell type to invade and discuss other approaches to modelling the invasion of GBM cell populations. Finally, we explore the dynamics of the model through conducting a phase plane analysis and studying the existence of travelling wave solutions.

In Chapter 3, we present information from image-localized biopsies that provides insight into the distribution and co-occurrence of EGFR and PDGFRA amplified sub-populations throughout GBM tumours in an initial cohort of patients. Using our mathematical model of interacting GBM sub-populations, we investigate the effects of different interaction assumptions, namely cooperative, competitive and neutral (no) interactions, on the population level occurrence of EGFR and PDGFRA amplified cells *in silico*. We study population levels found across the image-localized biopsy data from a cohort of patients and compare this to model outputs under these different interaction assumptions. We explore additional factors affecting the patterns observed in our simulations, such as selection advantages and phylogenetic ordering of mutations, which may also contribute to the levels of EGFR and PDGFRA amplified populations observed in biopsy data. Finally, we conduct a sensitivity analysis of our model and discuss our results and the insight they provide into the evolution of these biologically complex tumours.

We begin Chapter 4 by expanding the image-localised biopsy dataset to include data from additional patients and discussing the data in further detail. We provide a patient example, illustrating the heterogeneity in amplification of the EGFR and PDGFRA genes throughout this individual's tumour and present a summary of the dataset. We then discuss the need to introduce stochasticity into our model formulation, in order to reflect the variation in amplification lev-

els of EGFR and PDGFRA observed across the patient dataset. We do this by assuming that a subset of the model parameters have an associated probability distribution and explore the effect that this has on the variety of amplification patterns observed in a simulated cohort of GBMs. Following on from this, we create synthetic datasets and test whether we are able to infer the true model parameters and discuss a number of challenges relating to this. Finally, we infer the model parameters for the patient dataset in the hope that this will shed some light on the type of interactions occurring between EGFR and PDGFRA amplified sub-populations in GBMs and we discuss the possible implications of our findings.

The conclusions drawn from these investigations are summarised in Chapter 5, where we discuss the main findings of this work and suggest directions for future investigations.

Chapter 2

Modelling glioblastomas with multiple interacting sub-populations

This chapter begins by presenting a novel model of interacting sub-populations in glioblastomas. We briefly discuss different approaches to modelling the invasion of multiple GBM cell sub-populations, before presenting the derivation of the movement term included in our model. Following on from this, we study the spatially homogenous steady states of the model and present a phase plane analysis, before studying travelling wave solutions to the model.

2.1 A Model of Interacting EGFR and PDGFRA amplified sub-populations in Glioblastomas

Over recent years numerous different approaches have been taken to modelling the growth of GBM; such approaches include multiscale, lattice-based or stochastic models, with each having a focus on capturing particular properties of these complex tumours; a recent comprehensive review of mathematical approaches to modelling GBMs is given by Alfonso et al. in [3]. The approach we follow here, however, is inspired by the "Proliferation-Invasion" (PI) model, which takes the form of the well-known Fisher-KPP equation [115, 116, 117, 119, 120, 121]. The PI model is a minimal model of glioblastoma growth based on the simplified definition of cancer as "uncontrolled proliferation of cells with the ability to invade" [115]; two phenomena that GBMs are well-known to exhibit aggressively. The real power of the PI model lies in its simplicity as model parameters can be esti-

mated from patient MRI scans, allowing patient-specific predictions to be made and inform clinical practice [6, 45, 116, 130]. For this reason, we choose to adopt a similar minimal approach to modelling the growth of EGFR and PDGFRA amplified sub-populations in GBMs.

Our model, therefore, takes the form of an extended PI model to account for the growth of three genetically-distinct sub-populations defined as tumour cells with the genes encoding for EGFR (E), PDGFRA (P) and neither (N) protein amplified. The model is given by:

$$\frac{\partial E}{\partial t} = \nabla \cdot \left(D_E \left(1 - \frac{P+N}{K} \right) \nabla E + D_E \frac{E}{K} (\nabla P + \nabla N) \right) + f_E(E,P,N), \quad (2.1)$$

$$\frac{\partial P}{\partial t} = \nabla \cdot \left(D_P \left(1 - \frac{E+N}{K} \right) \nabla P + D_P \frac{P}{K} (\nabla E + \nabla N) \right) + f_P(E,P,N), \quad (2.2)$$

$$\frac{\partial N}{\partial t} = \nabla \cdot \left(D_N \left(1 - \frac{E+P}{K} \right) \nabla N + D_N \frac{N}{K} (\nabla E + \nabla P) \right) + f_N(E,P,N). \quad (2.3)$$

We consider this model on a one-dimensional cartesian domain, $x \in [0, L]$, with zero flux boundary conditions at $x = 0, L$, and the initial conditions given by,

$$E(\mathbf{x}, 0) = E_0(\mathbf{x}), \quad P(\mathbf{x}, 0) = P_0(\mathbf{x}) \quad \text{and} \quad N(\mathbf{x}, 0) = N_0(\mathbf{x}), \quad (2.4)$$

where E, P and N are the concentrations of each of the tumour cell sub-populations (cells/mm^3) and E_0, P_0 and N_0 are suitable functions defining their spatial distributions at time $t = 0$.

Similarly to the PI model, our model essentially consists of two terms to model the evolution of each cell sub-population. One of these models the uncontrolled proliferation of each tumour cell sub-population, given by the terms f_E, f_P and f_N, which take the form:

$$f_E(E,P,N) = \rho_E E \left(1 + \alpha_{PE} \frac{P}{K} \right) \left(1 - \frac{E+P+N}{K} \right), \quad (2.5)$$

$$f_P(E,P,N) = \rho_P P \left(1 + \alpha_{EP} \frac{E}{K} \right) \left(1 - \frac{E+P+N}{K} \right), \quad (2.6)$$

$$f_N(E,P,N) = \rho_N N \left(1 - \frac{E+P+N}{K} \right). \quad (2.7)$$

These proliferative terms include a joint logistic growth factor, where we assume there are plenty of other resources such as oxygen and nutrients available and the proliferation of each population is limited only by the availability of space, such that no net proliferation occurs once the *maximum carrying capacity* (K),

defined as the maximum number of cells that can fill a given volume, is reached. We derive an approximate value for K (as in [122]) by assuming that all sub-populations have cells of the same size with a radius of 10μm, yielding a volume of approximately $4.189 \times 10^3 \mu\text{m}^3$. Thus, we have a maximum carrying capacity of

$$K = \frac{1\,\text{cell}}{4.189 \times 10^3 \mu\text{m}^3} \left(\frac{10^3 \mu\text{m}}{1\text{mm}} \right)^3 = 2.39 \times 10^5 \frac{\text{cells}}{\text{mm}^3}. \tag{2.8}$$

The other factors in the above proliferation terms can be thought of as modified net proliferation rates, where the parameters ρ_E, ρ_P and ρ_N represent the intrinsic net proliferation rates of each population (1/year). The parameters α_{EP} and α_{PE} measure the effect of sub-population E on sub-population P and vice versa. For example, if $\alpha_{EP} > 0$ in Eq. (2.6), then the presence of the EGFR amplified (EGFRamp) population, E, promotes proliferation of the PDGFRA amplified (PDGFRAamp) population, P; this could be due to secretion of a growth factor that PDGFRAamp cells are sensitive to, for example. Alternatively, if $\alpha_{EP} < 0$, then the net proliferation of PDGFRAamp cells reduces as the density of the EGFRamp population increases and the PDGFRAamp cells are negatively affected. Furthermore, if $\alpha_{EP} = 0$, then sub-population E has no effect on the net proliferation of P. The parameter α_{PE} is defined analogously and we note that if both α_{EP} and α_{PE} are zero, then there are no additional interactions between the two populations, only competition for space. We define the types of interactions that can occur between EGFR and PDGFR amplified sub-populations in our model and summarise them in Table 2.1.

Table 2.1: The definitions of interactions that can occur between sub-populations E and P in our mathematical model as determined by the signs of α_{EP} and α_{PE}.

α_{EP}	α_{PE}	Interaction Type
0	0	Neutralism
< 0	0	Amensalism: E negatively affects P
0	< 0	Amensalism: P negatively affects E
< 0	< 0	Competition
> 0	0	Commensalism: E positively affects P
0	> 0	Commensalism: P positively affects E
> 0	> 0	Cooperation
> 0	< 0	Parasitism: of P on E
< 0	> 0	Parasitism: of E on P

Meanwhile, the remaining term on the right hand side of Eq.s (2.1)-(2.3) models the ability of each cell type to invade. Each of these terms model the net migration of each population as a form of diffusion, where the parameters D_E,

D_P and D_N represent their diffusion coefficients (mm^2/year). This non-linear form of diffusion is termed *cross-diffusion* and is characterised by a gradient of one population inducing a flux of the other [59]. We choose to incorporate this form of diffusion in our model since we expect the migration of tumour cell sub-populations to be affected by the presence of other tumour cells and we derive these terms following the volume-filling approach of Painter and Hillen [78] in Section 2.2.1. This approach is based on the assumption that a finite number of cells, the *maximum carrying capacity* (K), can fill a given volume and cells will continue to fill that space until this capacity is reached; we assume this number of cells, K, to be the same for each of the cell populations E, P and N, as previously derived.

Before moving on, we observe briefly that under the assumption that the EGFR and PDGFRA amplified sub-populations do not interact with one another, i.e. $\alpha_{EP} = \alpha_{PE} = 0$, and all three sub-populations diffuse and proliferate at the same rates, i.e. $D_E = D_P = D_N$ and $\rho_E = \rho_P = \rho_N$, then the system can be reduced to a single equation governing the total population of tumour cells, $T = E + P + N$; this equation is the well-known, clinically significant Proliferation-Invasion (PI) model that describes the evolution of a single homogeneous population of GBM tumour cells mentioned at the start of this section. Therefore, we note that our model is consistent with the PI model when used to model a single homogeneous population T.

In the following section, we discuss some alternative approaches to modelling the migration and invasion of multiple GBM tumour cell sub-populations, before proceeding to present the derivation of the cross-diffusion terms that arise in our model, given by Eq.s (2.1)-(2.3).

2.2 Modelling the invasion of GBM cell populations

Over recent years, many mathematical models have been proposed in an attempt to replicate patterns of GBM invasion observed *in vivo* and *in vitro*. In a recent review by Alfonso et al. [3], a wide variety of such models are discussed, taking a range of forms including cellular automaton, lattice-gas cellular automaton, cellular Potts models, partial differential equations, agent-based models and evolutionary game theory models; a list of relevant references can be found in [3]. Here, however, we focus on partial differential equation (PDE) models, as we wish to focus our attention on extending the patient-specific PI model framework

in this and future work.

Within the class of PDE models, several different approaches to modelling the spread of GBMs have been published. Perhaps the most well-known of these is the PI model by Swanson et al. [115, 116, 117, 118, 120, 121]. The PI model consists of a single equation governing the evolution of the tumour cell population, u, given by

$$\frac{\partial u}{\partial t} = \nabla \cdot (D \nabla u) + \rho u \left(1 - \frac{u}{K}\right); \tag{2.9}$$

a more in depth discussion of this model and its clinical relevance can be found in Section 1.2.1. We highlight here, however, the term modelling the net migration of the tumour cells, that is $\nabla \cdot (D \nabla u)$. This term models the movement of cells as isotropic diffusion, where the parameter D is either a constant diffusion coefficient [115, 116, 118] or is assigned different values in grey and white matter, allowing for a greater motility coefficient in the white matter of the brain, as observed biologically [117, 119, 120, 121]. In spite of the relative simplicity of the PI model, the patient-specific image driven nature of Swanson et al.'s approach has led to it being found to be prognostically significant in several instances [6, 45, 74].

Adding to this work, several groups have modelled cell migration as anisotropic diffusion by incorporating a diffusion tensor in their models to account for the preferential migration of tumour cells along white matter tracts [46, 76, 114]. A diffusion tensor image (DTI) can be used to map white matter tracts throughout the brain, since water molecules can move more freely along these fibres compared to perpendicular movement [2, 47, 114]. To this end, Jbabdi et al. [46] derive a diffusion tensor from a DTI scan of a healthy individual and model the migration of cells using a "Fickian" form of diffusion, i.e. $\nabla \cdot (D \nabla u)$, where D is a three-dimensional diffusion tensor, and the results were compared visually to real gliomas. Swan et al. [114] note, on the other hand, that the relation between the diffusion tensor and DTI data is not well justified in this work, along with others taking a similar approach [18, 52, 69]; in their work [46], Jbabdi et al. artificially increase the anisotropy of the DTI data by some unknown factor and it is also unclear how the measured water diffusion tensors are scaled to cell movement anisotropies. Alternatively, Swan et al. adopt a cell-based approach to connect the DTI to an effective tumour diffusion tensor D developed by Painter and Hillen [76]. The model of cell movement the authors use takes the form of the less well known "Fokker-Planck" diffusion, $\nabla\nabla : (Du)$, where the colon denotes

the contraction of two tensors [114], given by

$$\nabla\nabla : (Du) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}(D_{ij}u). \tag{2.10}$$

This form of diffusion has been derived in other biologically relevant contexts [76, 109] and we observe that upon expansion,

$$\nabla\nabla : (Du) = \nabla \cdot (D\nabla u) + \nabla \cdot ((\nabla^T D)u),$$

we obtain a Fickian diffusion term plus advection. The advective term in this model directs cell movement according to the spatially varying anisotropy, given by D, which results in increased migration of cells in the direction of white matter tracts, while slowing invasion in orthogonal directions [76]. The authors proceed to compare simulations using their anisotropic model to those produced using the PI model and found that in nine out of ten cases their model provided a better fit [114].

While the models discussed here all describe the spatial and temporal evolution of a single population of tumour cells, GBMs are known to consist of multiple genetically and phenotypically distinct subclones [82] and the question of how to model a population comprised of multiple distinct sub-populations remains. It may be reasonable to assume movement is independent of other cells if the populations are highly dispersed, for an example see reference [96]. On the other hand, for denser tissues, such as GBMs, contacts between cells of different populations are likely and they may compete with one another for space and resources or perhaps form a mutualistic or cooperative relationship. The nature of these interactions are modelled through reaction terms in our work and will be investigated in Chapters 3 and 4; we ask here, however, how simply the presence of other tumour sub-populations may affect how a given sub-population migrates and discuss some approaches to modelling this using PDEs.

In the PIHNA model, which was discussed briefly in Section 1.2.2, Swanson et al. [122] incorporate three distinct glioma cellular compartments (normoxic, hypoxic, and necrotic), along with a vascular compartment and diffusible angiogenic factors. In this model it is assumed that cells migrate by a Fickian form of diffusion, but also compete with other neighbouring cells for space; this is modelled by introducing a density-dependent diffusion coefficient, which effectively decreases to zero as the total cell density approaches some finite carrying capacity. Papadogiorgaki et al. [79] use this same form of diffusion in their model of

glioma invasion, which is an extension to the PIHNA model to include an extra compartment for hypoglycemic cells.

An alternative form of model of GBM growth, termed "reaction-cross-diffusion systems" [15], are characterised by a gradient of one population inducing a flux of the other. An example of such a model was derived by Gerlee et al. [32, 33] to analyse effects of the "go-or-grow" hypothesis on macroscopic tumour growth; this hypothesis says that glioma cells have the ability to switch between proliferative and migratory phenotypes in response to micro-environmental changes and local cell density [3]. The authors first propose an individual-based stochastic model, in which cells are either in proliferative or migratory states, and proceed to derive a continuum approximation in the form of a coupled system of reaction-cross-diffusion equations for the two species, given by

$$\frac{\partial p}{\partial t} = D_\alpha(1-p-m)\frac{\partial^2 p}{\partial x^2} + \alpha p(1-p-m) - (q_m+\mu)p + qp_m, \tag{2.11}$$

$$\frac{\partial m}{\partial t} = D_\nu\left((1-p)\frac{\partial^2 m}{\partial x^2} + m\frac{\partial^2 p}{\partial x^2}\right) - (q_p+\mu)m + q_m p, \tag{2.12}$$

where $p(x,t)$ and $m(x,t)$ denote the densities of proliferating and migratory cells, respectively. In this model, p cells proliferate at the rate α and the diffusion coefficient $D_\alpha = \alpha/2$ captures tumour expansion driven by this proliferation. Meanwhile, the diffusion coefficient, D_ν, comes from the random movement of the migratory, m, cells. The parameter μ represents the rate at which all tumour cells die through apoptosis and q_p is the rate at which m cells switch to become p cells and q_m is defined analogously.

The resulting migration terms in this model differ somewhat between the two phenotypes of cells; the proliferative cells diffuse with some density dependence, whereas the movement of migratory cells is dependent on the second derivative of both species [32], which is typically found in two species size exclusion processes [14]. This model has been found to admit travelling wave solutions, which have been analysed in detail in [33].

Such cross-diffusion systems have also arisen outside the field of GBM modelling in the context of chemotaxis models [77] and can be derived using the volume-filling approach detailed by Painter and Hillen [78], which we now demonstrate by deriving a model of multiple interacting sub-populations.

2.2.1 Derivation of a Multiple Species Cross-Diffusion Model

In this section, the derivation of a cross-diffusion model is presented. This model is derived based on the volume-filling approach of Painter and Hillen [78], which assumes that cells move based on the space available for them to move into and will continue to do so until the space becomes "full" and cells are tightly packed. Unlike most tissues that have closely regulated cell densities to prevent cells becoming tightly packed and the resulting depletion of resources and regions of necrosis (dead tissue), GBMs are characterised by such regions of high cell density and a necrotic tumour core [17]. Thus, in order to allow such tight cell packing to occur in our model of GBM evolution, we choose to derive the invasion terms in our model (Eq.s (2.1)-(2.3)) based on this volume-filling approach.

The derivation presented here also uses the approach of Othmer and Stevens [109], wherein a continuous-time discrete-space master equation is considered before arriving at a system of partial differential equations, upon reinterpreting space as a continuous variable. We note that both the approaches of Painter and Hillen [78] and Othmer and Stevens [109] are presented in the context of chemotaxis modelling, however we neglect chemotaxis terms in this case. We also neglect reaction and proliferation terms for simplicity, i.e. $f_E = f_P = f_N = 0$, and present the derivation in one spatial dimension, with the result being easily extended to include higher dimensions.

We begin by assuming that the tissue is composed of three different cell types, E, P and N, defined on a 1D lattice with uniform spacing h. We define $E_i(t)$ to be the density of cell type E at lattice point i at time t and define $P_i(t)$ and $N_i(t)$ similarly. We assume that cell movement occurs by individuals moving freely into neighbouring unoccupied space and that the change in cell density at a lattice point i, therefore, evolves according to the continuous-time discrete-space equation given by,

$$\frac{\partial E_i}{\partial t} = \mathcal{T}^+_{i-1,E} E_{i-1} + \mathcal{T}^-_{i+1,E} E_{i+1} - (\mathcal{T}^+_{i,E} + \mathcal{T}^-_{i,E}) E_i. \tag{2.13}$$

In the above equation, the $\mathcal{T}^{\pm}_{i,E}$ define the transition probabilities per unit time of a one-step jump of type E cells to lattice point $i \pm 1$. We make the assumption that the three cell types, E, P and N, move in the same way and obtain analogous expressions that govern populations P and N.

At this stage, the form of the transition probabilities plays an important role in the type of equation that we proceed to derive. In previous work by Painter &

Hillen [78], the authors assume that cells move chemo-sensitively and, therefore, that the transition probabilities depend on the gradient of an external chemotactic gradient, leading to the derivation of the "classical chemotaxis model" as in previous work by Patlak [87] and Keller & Segel [48]. In [78], the authors then proceed to derive a chemotaxis model for a single population of cells whilst incorporating volume-filling effects. This is the approach we follow here, however we neglect chemo-sensitive movement and include two additional cell populations. Therefore, we choose the transition probabilities to be of the form

$$\mathcal{T}_{i,E}^{\pm} = \alpha q(E_{i\pm1}, P_{i\pm1}, N_{i\pm1}), \tag{2.14}$$

where $q(\cdot)$ is the probability of cells finding space at their neighbouring locations and α is some constant. We note that in this instance we assume the probabilities depend only on the space jumped to and not the starting point; an example of some work where different assumptions on the form of the transitional probabilities are used is discussed later in this section. We also assume that the transition probabilities for populations E and P take the same form, but that the rate at which the transitions take place may differ between the populations. Thus, we assume $\mathcal{T}_{i,P}^{\pm} = \beta\mathcal{T}_{i,E}^{\pm}$ and $\mathcal{T}_{i,N}^{\pm} = \gamma\mathcal{T}_{i,E}^{\pm}$, where β and γ are some constants.

Assuming that there exists a finite number, K, of cells which can occupy each lattice site, we require q to satisfy the following conditions:

$$q(E, P, N) = 0, \quad \text{when} \quad E + P + N = K \tag{2.15}$$

and

$$q(E, P, N) \geq 0, \quad \text{for all } (E, P, N) \text{ such that } \quad 0 \leq E + P + N \leq K. \tag{2.16}$$

Next, substituting expression (2.14) into Eq. (2.13) and rearranging yields the continuous-time discrete-space master equation governing population E given by

$$\begin{aligned}\frac{\partial E_i}{\partial t} &= \alpha q(E_i, P_i, N_i)E_{i-1} + \alpha q(E_i, P_i, N_i)E_{i+1} \\ &\quad - \alpha(q(E_{i+1}, P_{i+1}, N_{i+1}) + q(E_{i-1}, P_{i-1}, N_{i-1}))E_i \\ &= \alpha q(E_i, P_i, N_i)(E_{i+1} - 2E_i + E_{i-1}) - \alpha E_i(q(E_{i+1}, P_{i+1}, N_{i+1}) \\ &\quad - 2q(E_i, P_i, N_i) + q(E_{i-1}, P_{i-1}, N_{i-1}))\end{aligned} \tag{2.17}$$

and we obtain analogous expressions for the P and N populations of cells.

Defining $x = ih$ and reinterpreting x as a continuous variable, we use Taylor's

Theorem to expand the terms inside each of the brackets on the right hand side (RHS) of Eq. (2.17), to find that

$$q(E_{i+1},P_{i+1},N_{i+1})-2q(E_i,P_i,N_i)+q(E_{i-1},P_{i-1},N_{i-1}) = h^2\frac{\partial^2 q(E_i,P_i,N_i)}{\partial x^2}+\mathcal{O}(h^4) \tag{2.18}$$

and

$$E_{i+1}-2E_i+E_{i-1} = h^2\frac{\partial^2 E_i}{\partial x^2}+\mathcal{O}(h^4). \tag{2.19}$$

As we change the spatial scale, h, the probability of jumping to a neighbouring location should also depend on that scale and we assume, therefore, that they satisfy

$$\mathcal{T}_h^{\pm} = \frac{\mu}{h^2}\mathcal{T}^{\pm},$$

for each population, where μ is a scaling constant. In this way,

$$\mathcal{T}_{h,E,i}^{\pm} = \frac{\mu\alpha}{h^2}q(E_i,P_i,N_i)$$

and upon updating the master equation accordingly, substituting in the above expressions and taking the limit as $h\to 0$, we find that

$$\frac{\partial E}{\partial t} = D_E q\frac{\partial^2 E}{\partial x^2} - D_E E\frac{\partial^2 q}{\partial x^2}, \tag{2.20}$$

assuming that $\mu\alpha = D_E$. Following the same process for the P and N populations of cells and assuming $\mu\alpha\beta = D_P$ and $\mu\alpha\gamma = D_N$, we find analogous governing equations.

By carefully expanding the derivatives on the RHS of Eq. (2.20), it can be written in divergence form as

$$\frac{\partial E}{\partial t} = \frac{\partial}{\partial x}\left(D_E q\frac{\partial E}{\partial x} - D_E E\left(q_E\frac{\partial E}{\partial x}+q_P\frac{\partial P}{\partial x}+q_N\frac{\partial N}{\partial x}\right)\right), \tag{2.21}$$

where $q_E = \partial q/\partial E$ and q_P and q_N are similarly defined. As mentioned previously, the derivation can easily be extended to two or three spatial dimensions by assuming a von Neumann neighbourhood, i.e. cells can move to their four and six nearest neighbours on square and cubic lattices, respectively. Thus, the general volume-filling model can be written as

$$\frac{\partial E}{\partial t} = \nabla \cdot (D_E q \nabla E - D_E E(q_E \nabla E + q_P \nabla P + q_N \nabla N)), \tag{2.22}$$
$$\frac{\partial P}{\partial t} = \nabla \cdot (D_P q \nabla P - D_P P(q_E \nabla E + q_P \nabla P + q_N \nabla N)), \tag{2.23}$$
$$\frac{\partial N}{\partial t} = \nabla \cdot (D_N q \nabla N - D_N N(q_E \nabla E + q_P \nabla P + q_N \nabla N)), \tag{2.24}$$

where ∇ is the appropriate differential operator, e.g. $\nabla = (\partial/\partial x, \partial/\partial y)$ in two spatial dimensions. We note that the derivation outlined above can be easily be extended to account for $M \in \mathbb{N}$ distinct populations to obtain a system of M coupled reaction-cross-diffusion equations, with the general form given by

$$\frac{\partial u^k}{\partial t} = \nabla \cdot \left(D_{u^k} q \nabla u^k - D_{u^k} u^k \sum_{j=1}^{M} q_{u^j} \nabla u^j \right), \quad \text{for } k = 1, ..., M, \tag{2.25}$$

where u^i and D_{u^i} denote the density and diffusion coefficient of sub-population i, respectively, $q = q(u^1, ..., u^M)$ and q_{u^i} denotes the derivative of the function q with respect to the variable u^i.

Finally, it remains to choose a suitable form of $q(\cdot)$, the function defining the probability of cells finding space at their neighbouring locations. In [78], the authors discuss possible forms for this function, suggesting various linear and non-linear options to incorporate various cellular processes. In this instance, however, we choose q to be given by

$$q(E, P, N) = 1 - \frac{E + P + N}{K}, \tag{2.26}$$

which simply states that the probability of a cell moving to a lattice point decreases linearly with the total density of cells at that point. We note that this satisfies conditions (2.15) and (2.16) and, upon substituting into Eq.s (2.22)-(2.24), we obtain the model equations:

$$\frac{\partial E}{\partial t} = \nabla \cdot \left(D_E \left(1 - \frac{P + N}{K} \right) \nabla E + D_E \frac{E}{K} (\nabla P + \nabla N) \right), \tag{2.27}$$
$$\frac{\partial P}{\partial t} = \nabla \cdot \left(D_P \left(1 - \frac{E + N}{K} \right) \nabla P + D_P \frac{P}{K} (\nabla E + \nabla N) \right), \tag{2.28}$$
$$\frac{\partial N}{\partial t} = \nabla \cdot \left(D_N \left(1 - \frac{E + P}{K} \right) \nabla N + D_N \frac{N}{K} (\nabla E + \nabla P) \right). \tag{2.29}$$

We mentioned previously that we wanted our model to take the form of an

extended PI model for three populations of cells. While this derived movement equation looks much more complicated than the type of movement modelled by the PI model, that is Fickian diffusion, we note that in the analogous single species model for cell population u^1 (i.e. Eq. (2.25) with $N = 1$), choosing q to be of an analogous linear form, that is

$$q(u^1) = 1 - \frac{u^1}{K}, \tag{2.30}$$

yields the standard Fickian diffusion equation. Thus, the terms used to model the invasion of tumour cells in our model provide a natural extension to the cell movement term employed in the PI model in a three-species setting.

We note that other cross-diffusion models have also been proposed. Ostrander [75] derived a cross-diffusion model for two interacting populations using a similar approach to Painter and Hillen [78], but argued that the transition probabilities, $\mathcal{T}^{\pm}_{i,u}$ and $\mathcal{T}^{\pm}_{i,v}$, should depend on the desire of each species to leave a lattice point i along with the favourableness of moving to the neighbouring point $i \pm 1$. Thus, the transition probabilities used depend on the population densities at the lattice point i as well as $i \pm 1$ and do not satisfy the same conditions that we imposed on the function q in our derivation, which was based on a volume-filling assumption. The author also allowed the transition probabilities to be composed of different functions for each species [75]; in the above derivation we assumed that the three populations move in the same way and that the transition probabilities, therefore, differ only by some scaling constant, that is $\mathcal{T}^{\pm}_{i,v} = \beta \mathcal{T}^{\pm}_{i,u}$. However, this could easily be altered in our derivation if biological evidence becomes available to suggest otherwise.

A second model incorporating cross-diffusion terms arose in a paper by Painter [77], where a chemotaxis model describing a heterogeneous tissue comprising two distinct populations of motile cells was derived. The aim of this work was to explore the capacity for differential chemotaxis to drive sorting/patterning of a heterogeneous tissue. The model derived in this paper consisted of a coupled system of PDEs, which included cross-diffusion terms analogous to those derived in Eq.s (2.27)–(2.29) as well as the relevant terms corresponding to chemotactic movement. This model also included another term describing cell movement, which arose by assuming cells not only move unimpeded into neighbouring unoccupied space, but also by an "active" cell (i.e., one that pulls forward) displacing a "passive" cell (i.e., one that is pulled back) and moving into neighbouring occupied space via a process of "location-swapping" [77]. Following this approach yields

similar equations to those derived in Eq.s (2.27)–(2.29), but with an additional term corresponding to the movement by cells "swapping locations". This approach could be another suitable candidate to describe the movement of multiple sub-populations in GBMs, however, it increases model complexity and introduces more parameters that will need to be estimated. Thus, we leave the incorporation of "location swapping" terms into our model as a consideration for future work.

2.3 Travelling waves in the model of interacting sub-populations

We are interested in the patterns of invasion exhibited in glioblastomas and the distribution of EGFR and PDGFRA amplified tumour sub-populations within them; for example, whether they co-exist or occupy distinct tumour regions or whether one cell-type is typically present at the invading front and the other in the tumour core or both invade simultaneously. Therefore, in this section we study the behaviour of our model in more detail and the effect that different types of interactions between the amplified sub-populations has on the distributions and patterns of co-occurrence of cell-types that we observe. We begin by presenting a phase plane analysis, followed by studying the travelling waves exhibited by our model.

2.3.1 Phase Plane Analysis

To find the spatially homogeneous steady states of the model, we set the functions f_E, f_P and f_N, defined by equations (2.5)-(2.7), equal to zero. Thus, we find the following spatially homogeneous steady states:

$$(\bar{E}, \bar{P}, \bar{N}) = \quad (0,0,0), \quad (-K/\alpha_{EP}, -K/\alpha_{PE}, 0), \quad (K,0,0), \quad (0,K,0), \quad (0,0,K),$$

and the continuum of co-existence steady states $(\bar{E}, \bar{P}, \bar{N}) = (E^*, P^*, N^*)$, where $E^* + P^* + N^* = K$ and $0 < E^*, P^*, N^* < K$. We note that the steady state $(\bar{E}, \bar{P}, \bar{N}) = (-K/\alpha_{EP}, -K/\alpha_{PE}, 0)$ only exists and is biologically relevant when populations E and P are interacting with one another with sufficiently strong competition, that is $\alpha_{EP}, \alpha_{PE} < 0$ such that $1 + 1/\alpha_{EP} + 1/\alpha_{PE} \geq 0$, in order to ensure the conditions $E, P > 0$ and $E + P \leq K$ are satisfied.

First we consider the special case where only the interacting EGFR and PDGFRA amplified sub-populations are present by setting $N \equiv 0$ throughout

the following analysis. We note that the steady states of the reduced system are

$$(\bar{E},\bar{P}) = \quad (0,0), \quad (-K/\alpha_{EP}, -K/\alpha_{PE}), \quad (K,0), \quad (0,K),$$

and a continuum of co-existence steady states $(\bar{E},\bar{P}) = (E^*,P^*)$, where $E^*+P^* = K$ and $0 < E^*, P^* < K$.

In this case, the Jacobian of the spatially homogeneous system is given by,

$$J(E,P) = \begin{pmatrix} \frac{\partial f_E}{\partial E} & \frac{\partial f_E}{\partial P} \\ \frac{\partial f_P}{\partial E} & \frac{\partial f_P}{\partial P} \end{pmatrix} \tag{2.31}$$

where,

$$\frac{\partial f_E}{\partial E} = \rho_E \left(1 + \alpha_{PE}\frac{P}{K}\right)\left(1 - \frac{2E+P}{K}\right), \tag{2.32}$$

$$\frac{\partial f_E}{\partial P} = \alpha_{PE}\rho_E \frac{E}{K}\left(1 - \frac{E+P}{K}\right) - \rho_E \frac{E}{K}\left(1 + \alpha_{PE}\frac{P}{K}\right), \tag{2.33}$$

$$\frac{\partial f_P}{\partial E} = \alpha_{EP}\rho_P \frac{P}{K}\left(1 - \frac{E+P}{K}\right) - \rho_P \frac{P}{K}\left(1 + \alpha_{EP}\frac{E}{K}\right), \tag{2.34}$$

$$\frac{\partial f_P}{\partial P} = \rho_P \left(1 + \alpha_{EP}\frac{E}{K}\right)\left(1 - \frac{E+2P}{K}\right). \tag{2.35}$$

The zero steady state, $(\bar{E},\bar{P}) = (0,0)$, is unstable since the eigenvalues of matrix (2.31) are $\lambda_1 = \rho_E$ and $\lambda_2 = \rho_P$ and $\rho_{E,P} > 0$. At the competition co-existence steady state, $(\bar{E},\bar{P}) = (-K/\alpha_{EP}, -K/\alpha_{PE})$, we have

$$\mathbf{J}(\bar{E},\bar{P}) = \begin{pmatrix} 0 & -\frac{\alpha_{PE}}{\alpha_{EP}}\rho_E\left(1 + \frac{1}{\alpha_{EP}} + \frac{1}{\alpha_{PE}}\right) \\ -\frac{\alpha_{EP}}{\alpha_{PE}}\rho_P\left(1 + \frac{1}{\alpha_{EP}} + \frac{1}{\alpha_{PE}}\right) & 0 \end{pmatrix},$$

with eigenvalues given by

$$\lambda_{1,2} = \pm\sqrt{\rho_E\rho_P}\left(1 + \frac{1}{\alpha_{EP}} + \frac{1}{\alpha_{PE}}\right). \tag{2.36}$$

These are both real and of opposite sign, thus, the spatially homogeneous steady state $(\bar{E},\bar{P}) = (-K/\alpha_{EP}, -K/\alpha_{PE})$ is a saddle node when it exists (i.e. $\alpha_{EP}, \alpha_{PE} < 0$ and $1 + 1/\alpha_{EP} + 1/\alpha_{PE} \geq 0$).

The remaining steady states, namely $(\bar{E},\bar{P}) = (0,K), (K,O)$ and (E^*,P^*) where $E^* + P^* = K$ and $0 < E^*, P^* < K$, are all non-hyperbolic since the Jacobian matrix (defined by (2.31)-(2.35)) has a zero eigenvalue in each case. For

the two steady states where one population is extinct and the other is at carrying capacity, $(0, K)$ and $(K, 0)$, these are $\lambda_1 = 0$ and $\lambda_2 = -\rho_P$ or $-\rho_E$ respectively. These steady states each have a stable one dimensional manifold associated with a negative real eigenvalue of the Jacobian. Meanwhile, the sign of the non-identically-zero eigenvalue associated with the continuum of coexistence states, $(\bar{E}, \bar{P}) = (E^*, P^*)$ where $0 < E^*, P^* < K$, depends on the model parameters and the values of E^* and P^* and is given by

$$\lambda = -\rho_E \frac{E^*}{K}\left(1 + \alpha_{PE}\frac{P^*}{K}\right) - \rho_P \frac{P^*}{K}\left(1 + \alpha_{EP}\frac{E^*}{K}\right). \tag{2.37}$$

We note that this eigenvalue is real for each (E^*, P^*) pair and may be positive and, thus, unstable under certain parameter regimes where at least one of α_{EP} and α_{PE} are negative; we note that such possible parameter regimes are limited to the interaction types of amensalism, cooperation and parasitism, as defined in Table 2.1.

Plots of the phase planes of the spatially homogeneous model given by

$$\frac{dE}{dt} = f_E \quad \text{and} \quad \frac{dP}{dt} = f_P, \tag{2.38}$$

where f_E and f_P are defined by (2.5) and (2.6), respectively, with $N = 0$ are shown in Fig. 2.1 for example sets of parameters over the region of biologically relevant solutions, that is $0 \leq E, P \leq K$ and $E + P \leq K$ so that only concentrations of populations that are non-negative and do not exceed that carrying capacity are considered. In each plot, the model parameters ρ_E, ρ_P and K are kept the same and the parameters α_{EP} and α_{PE} are changed to illustrate how the phase space plots change according to the type of interaction between populations E and P. The interaction types shown are defined in Table 2.1 and are as follows: (a) neutralism; (b) amensalism, where population E negatively affects population P; (c) competition; (d) commensalism, where E positively affects P; (e) cooperation; and, finally (f) parasitism of E on P. We note that not all of the interactions given in Table 2.1 are shown in these plots since they are analogous to cases already shown, e.g. amensalism where P negatively affects E is analogous to the case where E negatively affects P. Trajectories of the model are also shown for a range of initial conditions, $E(0) = E_0$ and $P(0) = P_0$.

We observe that the cases of neutralism, commensalism and cooperation in Fig.s 2.1a, 2.1d and 2.1e look qualitatively similar, with the trajectories of populations E and P monotonically increasing until a steady state (E^*, P^*) along the line $E^* + P^* = K$ is reached. The position along this line and, thus, proportion

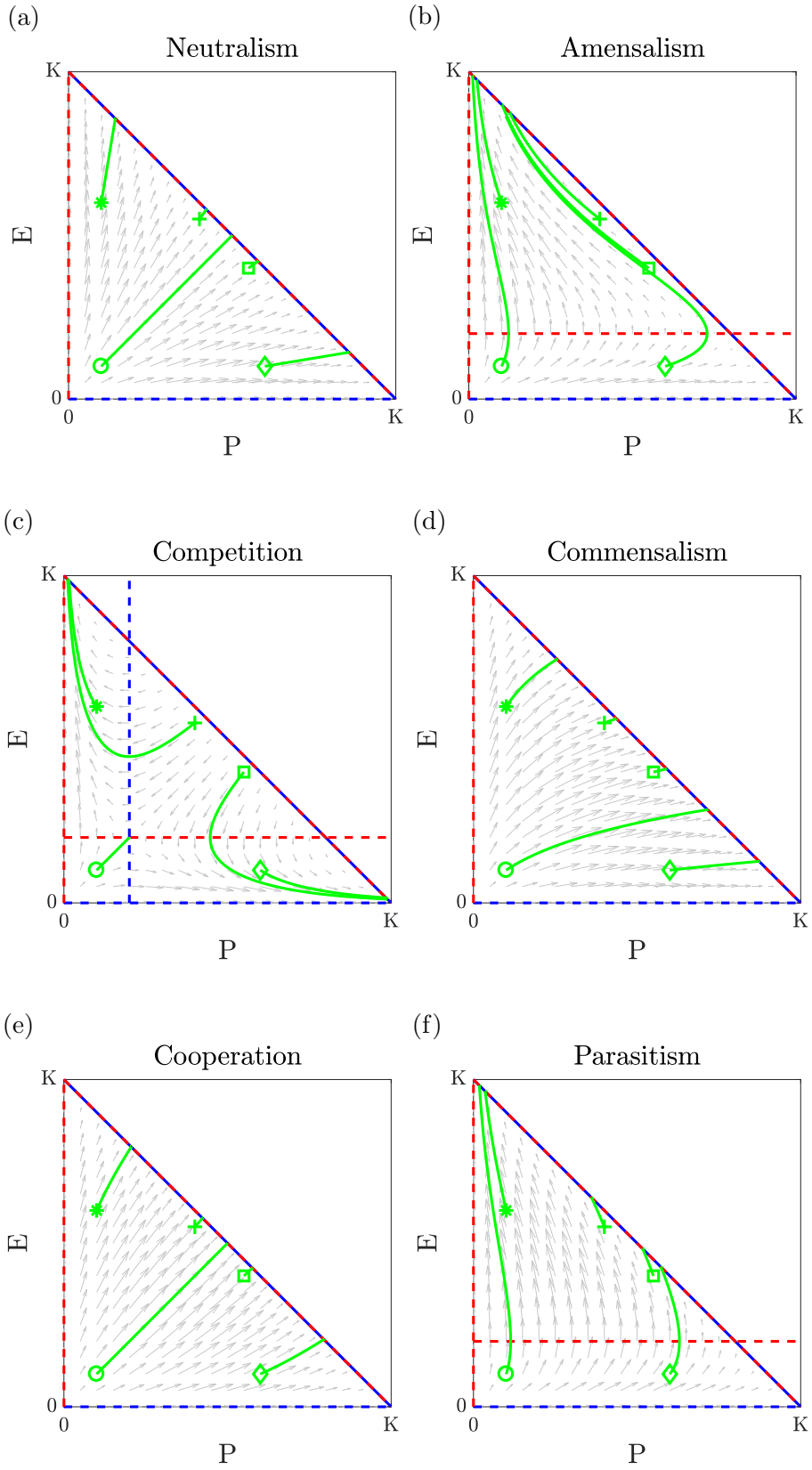

(a)
Neutralism
(b)
Amensalism
(c)
Competition
(d)
Commensalism
(e)
Cooperation
(f)
Parasitism
K
0
E
P

Figure 2.1

Figure 2.1: (See previous page for figure.) Plots showing E- and P-nullclines (blue and red dashed lines, respectively) and trajectories (green curves) of the model given by (2.38) (with $N = 0$) in $E - P$ phase space under different interaction types: (a) Neutralism, $\alpha_{EP} = 0$, $\alpha_{PE} = 0$; (b) Amensalism, $\alpha_{EP} = -5$, $\alpha_{PE} = 0$; (c) Competition, $\alpha_{EP} = -5$, $\alpha_{PE} = -5$; (d) Commensalism, $\alpha_{EP} = 5$, $\alpha_{PE} = 0$; (e) Cooperation, $\alpha_{EP} = 5$, $\alpha_{PE} = 5$; (e) Parasitism, $\alpha_{EP} = -5$, $\alpha_{PE} = 5$. Other model parameters used are $\rho_E = \rho_P = 15\text{years}^{-1}$ and $K = 2.39 \times 10^5\text{cells/mm}^3$. The initial conditions of the trajectories in each plot are: $*$ $E_0 = 0.6K$, $P_0 = 0.1K$; $\square$ $E_0 = 0.4K$, $P_0 = 0.55K$; $\diamond$ $E_0 = 0.1K$, $P_0 = 0.6K$; $\circ$ $E_0 = 0.1K$, $P_0 = 0.1K$; $+$ $E_0 = 0.55K$, $P_0 = 0.4K$. Simulations were produced using a Forward-Euler time stepping scheme for $t = 0$ to 2 years and a time-step of 0.0005 years.

of the populations E and P in the final state depend on the initial condition and type of interaction used in that simulation. We note that the proliferation rates ρ_E and ρ_P also affect this, but are kept constant in these examples; an example where these parameters are varied will be provided later.

Unlike the previous three cases where the model trajectories monotonically increase in E and P, this is not the case for the amensalism, competition and parasitism interaction types. In each of these scenarios at least one of α_{EP} and α_{PE} are negative resulting in regions of phase space where $f_E < 0$, $f_P < 0$ or both. These regions are divided by nullclines, shown by the blue and red dashed lines in Fig. 2.1. For example, in the amensalism case, the trajectories with initial conditions $\circ$ and $\diamond$ begin with both E and P increasing in a region where both $f_E, f_P > 0$ before crossing the P-nullcline into the region where the E population is large enough so that f_P becomes negative and P decreases. Furthermore, we previously noted that in each of these interaction scenarios it is possible for the non-zero eigenvalue, given by Eq. (2.37), associated with the continuum of steady states (E^*, P^*), such that $E^* + P^* = K$, to be positive and therefore unstable in that direction. We observe that there are regions where this is the case and find that the eigenvalue is positive for $0.276K < E^* < 0.724K$ and $0.053K < E^* < 0.947$ in the amensalism and competition examples given in Fig.s 2.1b and 2.1c, respectively.

Finally, we observe that the only trajectory where a steady state (E^*, P^*) with $E^* + P^* = K$ is not reached in these examples is in the competition case with initial condition $\circ$, $E_0 = P_0 = 0.1K$; in this case we see that the trajectory instead approaches the competition co-existence steady state, $(\bar{E}, \bar{P}) = (-K/\alpha_{EP}, -K/\alpha_{PE})$, along its stable manifold. This steady state is a saddle node and when it exists, as it does in this example (since $1 + 1/\alpha_{EP} + 1/\alpha_{PE} \geq 0$

and both α_{EP} and α_{PE} are negative), a separatrix divides phase space into regions separating different modes of behaviour of the dynamical system. Although the separatrix is not explicitly shown in Fig. 2.1c, we can see the difference between the trajectories initialised in different regions; for example, the trajectory with initial condition + approaches a steady state with population E high and P low, whereas the trajectory with □ initial condition approaches a steady state with the opposite. A trajectory will only approach a saddle node following a path along the separatrix, which we see is the case for the trajectory with initial condition ∘ in Fig. 2.1c. Separatrices are often difficult to find, but in this case can be found easily.

To find the equation of the separatrix, we look for solutions to the model that pass through the saddle node. We note that there are two solutions that fulfill this criteria: those that travel along the stable manifold of the saddle node, or the separatrix, towards the saddle node; and those that the travel away from the saddle node along its unstable manifold. We know that the stable manifold is tangent to the eigenvector, $\mathbf{v}_1$, associated with the negative eigenvalue at the saddle node, while the unstable manifold is tangent to the eigenvector, $\mathbf{v}_2$, associated with the positive eigenvalue. These are given by,

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ \frac{\alpha_{PE}}{\alpha_{EP}}\sqrt{\frac{\rho_E}{\rho_P}} \end{pmatrix}$$

and

$$\mathbf{v}_2 = \begin{pmatrix} 1 \\ -\frac{\alpha_{PE}}{\alpha_{EP}}\sqrt{\frac{\rho_E}{\rho_P}} \end{pmatrix}.$$

Therefore, we know that the separatrix will have a positive gradient at the saddle node, whereas the gradient of the unstable manifold will be negative.

Dividing the first equation of system (2.38) by the second, we find

$$\begin{aligned} \frac{\dot{E}}{\dot{P}} &= \frac{f_E(E,P,0)}{f_P(E,P,0)} \\ &= \frac{\rho_E(1+\alpha_{PE}P/K)}{\rho_E(1+\alpha_{EP}E/K)}, \end{aligned} \tag{2.39}$$

where $\dot{E}$ notation denotes differentiation with respect to t. Rearranging, we find that

$$\frac{1}{\rho_E}\left(\frac{\dot{E}}{E} + \frac{\alpha_{EP}}{K}\dot{E}\right) = \frac{1}{\rho_P}\left(\frac{\dot{P}}{P} + \frac{\alpha_{PE}}{K}\dot{P}\right). \tag{2.40}$$

Integrating with respect to t and rearranging, we find that

$$\frac{1}{\rho_E}\left(\ln E + \frac{\alpha_{EP}}{K}E\right) - \frac{1}{\rho_P}\left(\ln P + \frac{\alpha_{PE}}{K}P\right) = C, \tag{2.41}$$

where C is a constant. For different values of C, this equation gives the trajectories of the system in E-P space and, since trajectories along the separatrix must approach the saddle node, we can find the equation of the separatrix by requiring that

$$\lim_{t\to\infty} E(t) = \frac{-K}{\alpha_{EP}} \quad \text{and} \quad \lim_{t\to\infty} P(t) = \frac{-K}{\alpha_{PE}} \tag{2.42}$$

to find the value of C in Eq. 2.41 that corresponds to the separatrix. Thus, we find the separatrix is given by,

$$\frac{1}{\rho_E}\left(\ln\left(\frac{-\alpha_{EP}E}{K}\right) + \frac{\alpha_{EP}}{K}E + 1\right) - \frac{1}{\rho_P}\left(\ln\left(\frac{-\alpha_{PE}P}{K}\right) + \frac{\alpha_{PE}}{K}P + 1\right) = 0. \tag{2.43}$$

We note this equation is only defined when α_{EP} and α_{PE} are negative, which corresponds to the competition case and the existence of the competition coexistence state, $(\bar{E}, \bar{P}) = (-K/\alpha_{EP}, -K/\alpha_{PE})$.

Figure 2.2 shows the phase portrait of the dynamical system with a competitive interaction type for an example parameter set (as in Fig. 2.1c) with the solution curves of Eq. (2.43) shown in light blue. From this we see that there are, indeed two solution trajectories that pass through the saddle node given by Eq. (2.43): one defining the separatrix which corresponds to the stable one-dimensional manifold of the saddle node, along which solutions approach the saddle node; and the other corresponding to the one-dimensional unstable manifold, along which solutions move away from the saddle node. Since we know that the separatrix has a positive gradient at the saddle node, from Fig. 2.2, we see that the separatrix is the straight line $E = P$ for this example set of parameters, while the other pale blue curve given by solving Eq.(2.43) defines the unstable manifold. It is also clear from the phase portrait which curve represents the separatrix (stable manifold) and which the unstable manifold, as we see trajectories close to the separatrix moving towards the saddle node and then moving away as they become close to the unstable manifold.

Showing the separatrix explicitly in Fig. 2.2 clearly divides phase space into regions where trajectories behave differently; for example, the two trajectories with initial conditions lying to the left of the separatrix (the straight line $E = P$ in this case), $+$ and $*$, approach steady states with population E high and P low, whereas those lying to the right, $\square$ and $\diamond$, approach steady states with the

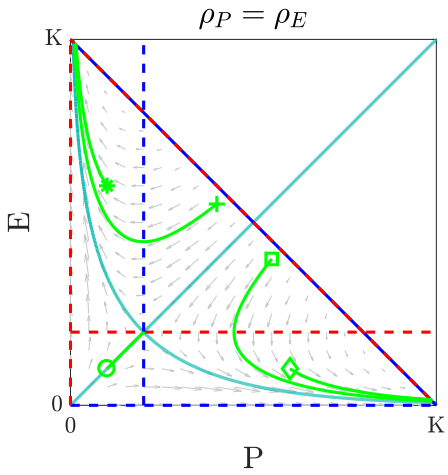


Figure 2.2: Phase portrait of the model given by (2.38) (with $N = 0$) in $E - P$ phase space under a competition interaction type, $\alpha_{EP} = -5$ and $\alpha_{PE} = -5$ (as in Fig. 2.1c) with solutions to Eq.(2.43), defining the stable and unstable manifolds, shown as light blue curves. All other information is the same as in Fig. 2.1c.

opposite. Since the trajectory with initial condition $\circ$ lies on the separatrix, we see this trajectory approaching the saddle node steady state along the separatrix, as expected.

Fig. 2.2 shows the special case of a phase portrait with a competitive interaction type where the parameters for each population are the same, that is $\alpha_{EP} = \alpha_{PE}$ and $\rho_E = \rho_P$. Since Eq. (2.43) defining the separatrix depends on these parameters, the separatrix curve may change as we vary these parameters, some examples of which are shown in Fig. 2.3. In the first two of these, Fig.s 2.3a and 2.3b, the two populations have equal proliferation rates ($\rho_E = \rho_P$) and different interaction parameters ($\alpha_{EP} \neq \alpha_{PE}$). In this case, the straight line

$$E = \frac{\alpha_{PE}}{\alpha_{EP}} P \tag{2.44}$$

solves Eq. (2.43) and defines the separatrix. Fig. 2.3a shows an example phase portrait with a competitive interaction type where population P is more competitive than population E, that is $\alpha_{PE} < \alpha_{EP} < 0$. In this example we see the separatrix shift so that more trajectories reach a steady state with P high and E low compared to when the competition parameters were equal, with four of the five example trajectories shown (green curves) now approaching a steady state of this type compared to only two out of five (see Figs 2.3a and 2.2, respectively). Similarly, when the competitive ability of P is reduced so that $\alpha_{EP} < \alpha_{PE} < 0$, we see in Fig. 2.3b that more trajectories approach steady states with E high and P low. This intuitively makes sense, since we expect a higher competitive abil-

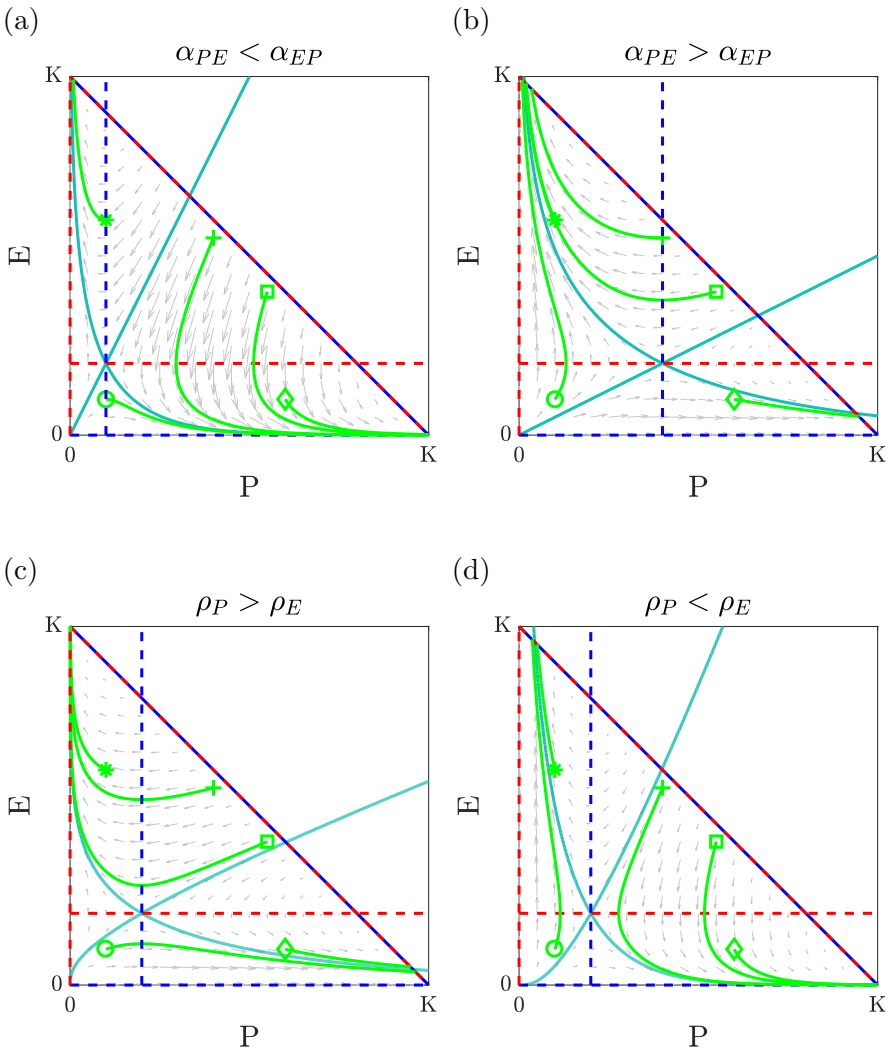


Figure 2.3: Plots showing the separatrix and unstable manifold (light blue lines) and example trajectories (green curves) of the model given by (2.38) (with $N = 0$) in $E - P$ phase space under a competition interaction type. In each plot $\rho_E = 15\text{years}^{-1}$ and $\alpha_{EP} = -5$, while the competitiveness and proliferative ability of population P is varied by changing the values of α_{PE} and ρ_P. These take the values (a) $\alpha_{PE} = 2\alpha_{EP}$ and $\rho_P = \rho_E$, (b) $\alpha_{PE} = \alpha_{EP}/2$ and $\rho_P = \rho_E$, (c) $\alpha_{PE} = \alpha_{EP}$ and $\rho_P = 3\rho_E$, and (d) $\alpha_{PE} = \alpha_{EP}$ and $\rho_P = \rho_E/3$. E- and P-nullclines are shown (blue and red dashed lines, respectively). All other information used to produce these plots are the same as in Fig. 2.1.

ity to be advantageous for a population over a less competitive one, resulting in the system approaching a steady state comprising more of the more competitive population.

It may also be reasonable to expect a higher proliferative ability to have a similar effect, however in Figs 2.3c and 2.3d we see that in the competitive case this is not necessarily the case. In these two examples the two populations are both equally competitive (α_{PE} and α_{EP} both negative and equal to one another),

while their relative proliferative ability is varied. Fig. 2.3c shows an example phase portrait where $\rho_P > \rho_E$, that is population P has a proliferative advantage over population E. In this case we see that only two of the five example trajectories shown approach steady states with P high and E low, with the remaining three approaching a steady state comprised mostly of population E. Of particular interest from these is the trajectory with initial condition $\square$. This trajectory begins with a higher proportion of the more proliferative population P present than the less proliferative population E, yet approaches a steady state comprised of only population E. This may seem counter intuitive, but upon examining the model equations (given by Eq.s (2.5), (2.6) (with $N \equiv 0$) and (2.38)) we see that dE/dt and dP/dt are both negative when E and P are large enough (but still such that $E + P < K$) in a competitive scenario due to the $(1 + \alpha_{EP}E/K)$ and $(1 + \alpha_{PE}P/K)$ terms. Thus, increasing the proliferation rate of population P makes dP/dt more negative in this instance so that this population decreases in size at a faster rate and eventually results in E becoming the dominant population, i.e. E comprises most of the final steady state of the system. We see a similar effect occurring with the trajectory with initial condition + in Fig. 2.3d when the proliferative ability of P is decreased so that E is the more proliferative population. These two examples illustrate that a proliferative advantage is not always advantageous for a tumour cell population and there are scenarios where the less aggressive tumour cell population will become the dominant one.

We briefly note that, in general, Eq. (2.43) is not defined when E and P are zero due to the logarithmic terms. However, we have seen that in the case where the proliferation rates, ρ_E and ρ_P, are equal, then Eq. (2.43) is solved by the straight line given by Eq. (2.44). Thus, the separatrix is exists when $E = P = 0$ in this instance. In the more general case when $\rho_E \neq \rho_P$, this not true and the separatrix is undefined at $E = P = 0$. From Fig.s 2.3c and 2.3d, however, we see that through solving Eq. (2.43) using MatLab and plotting the solution close to zero, the separatrix approaches the point $(E, P) = (0, 0)$.

Returning to the full system of equations with the non-amplified population of tumour cells, N, present, we can conduct a stability analysis of the spatially homogeneous steady states of the full model, given at the beginning of this section, in the same way. In this case, the Jacobian of the spatially homogeneous system is a 3×3 matrix defined analogously to that given by the 2×2 matrix (2.31) for the case with only two populations present. The eigenvalues of this matrix are calculated at each of the steady states to determine stability and are summarized as follows:

- $(\bar{E}, \bar{P}, \bar{N}) = (0, 0, 0)$ is an unstable steady state, since all eigenvalues are positive; $\lambda_1 = \rho_E$, $\lambda_2 = \rho_P$ and $\lambda_3 = \rho_N$.

- $(\bar{E}, \bar{P}, \bar{N}) = (-K/\alpha_{EP}, -K/\alpha_{PE}, 0)$ is an unstable steady state, with two positive and one negative eigenvalues; $\lambda_{1,2} = \pm\sqrt{\rho_E \rho_P}(1 + 1/\alpha_{EP} + 1/\alpha_{PE})$.

- $(\bar{E}, \bar{P}, \bar{N}) = (K, 0, 0)$, $(0, K, 0)$ and $(0, 0, K)$ are non-hyperbolic steady states with one negative eigenvalue, $\lambda_1 = -\rho_E$, $-\rho_P$ and $-\rho_N$, respectively, and the other two equal to zero.

- The continuum of steady states $(\bar{E}, \bar{P}, \bar{N}) = (E^*, P^*, N^*)$, where $E^* + P^* + N^* = K$ and $0 < E^*, P^*, N^* < K$ have two zero eigenvalues and the third is given by

$$\lambda_3 = -\rho_E \frac{E^*}{K}\left(1 + \alpha_{PE}\frac{P^*}{K}\right) - \rho_P \frac{P^*}{K}\left(1 + \alpha_{EP}\frac{E^*}{K}\right) - \rho_N \frac{N^*}{K}.$$

  This eigenvalue can be positive and, thus, a given steady state (E^*, P^*, N^*) can be unstable under certain parameter regimes where at least one of α_{EP} and α_{PE} are negative.

Figure 2.4 shows example phase portraits of the spatially homogeneous three population model for cooperative and competitive interaction types. In the cooperative case, Fig. 2.4a, each population increases and all model trajectories approach a steady state on the surface $E + P + N = K$ comprised of different proportions of each population depending on the initial conditions; this behaviour is similar to that observed in the two population case. In the competitive example, Fig. 2.4b, all trajectories, again, approach a steady state on the surface $E + P + N = K$. The trajectories with initial conditions $*$, $\diamond$ and $\circ$ each begin with a high proportion of one population present and approach a steady state comprised mostly of that same population. The other two trajectories behave slightly differently due to the competitive interactions. Recalling that the separatrix is the line $E = P$ when $\alpha_{EP} = \alpha_{PE} < 0$ and $\rho_E = \rho_P$ in the absence of N cells and since $E_0 = P_0$ for the $+$ and $\triangle$ initial conditions, each trajectory would move along the separatrix towards the competition co-existence steady state $((\bar{E}, \bar{P}) = (-K/\alpha_{EP}, -K/\alpha_{PE}))$ if $N_0 = 0$ as found previously (see Fig. 2.2). Instead, $N_0 > 0$ in this case so that as E and P approach $-K/\alpha_{EP}$ and $-K/\alpha_{PE}$, the N population increases, such that the system moves towards a steady state comprising mostly N cells.

Thus, we find that the behaviour of the spatially homogeneous model when non-amplified (N) tumour cells are present is similar to when this population is

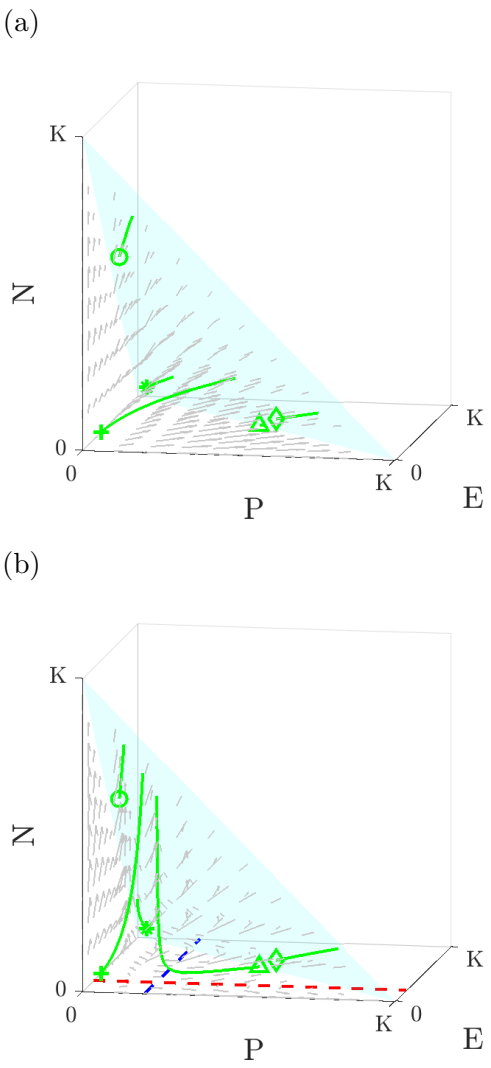


Figure 2.4: Phase portraits of the model given by $dE/dt = f_E$, $dP/dt = f_P$ and $dN/dt = f_N$ in $E - P - N$ phase space under (a) cooperation ($\alpha_{EP} = \alpha_{PE} = 5$) and (b) competition ($\alpha_{EP} = \alpha_{PE} = -5$) interaction types. In each plot $\rho_E = \rho_P = \rho_N = 15\text{years}^{-1}$. Model trajectories are plotted as green curves with initial conditions: $*$ $E_0 = 0.6K$, $P_0 = 0.1K$, $N_0 = 0.1K$; $\triangle$ $E_0 = 0.48K$, $P_0 = 0.48K$, $N_0 = 0.01K$; $\diamond$ $E_0 = 0.1K$, $P_0 = 0.6K$, $N_0 = 0.1K$; $\circ$ $E_0 = 0.1K$, $P_0 = 0.1K$, $N_0 = 0.6K$; $+$ $E_0 = 0.05K$, $P_0 = 0.05K$, $N_0 = 0.05K$. The nullcine $E + P + N = K$ is shown in cyan and the nullclines $E = -K/\alpha_{EP}$ and $P = -K/\alpha_{PE}$ are shown (red and blue dashed lines, respectively) projected onto the $E-P$ plane. Simulations were produced using a Forward-Euler time stepping scheme for $t = 0$ to 2 years and a time-step of 0.0005 years.

absent. This is because the non-amplified population of cells does not directly interact with the two amplified populations, only passively through competition for resources in the logistic growth term, whereas the more interesting model dynamics arise through interactions between the E and P populations of cells.

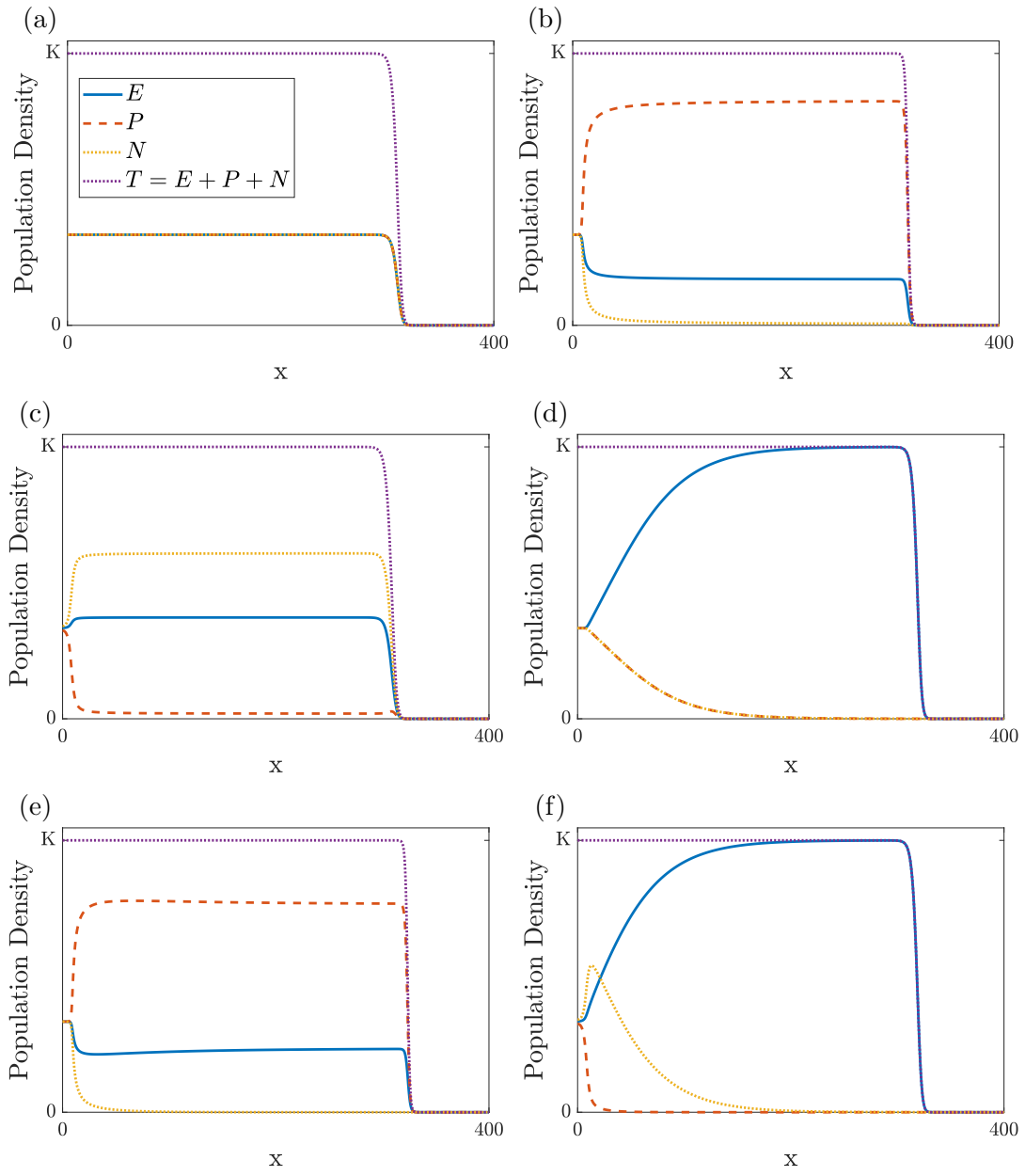


Figure 2.5: Plots showing solutions to the model converging to a range of travelling waves under different example parameter regimes on a one-dimensional domain, $x \in [0, 400]$mm. In each plot $D_E = D_P = D_N = 20$ mm^{2}year^{-1} and $\rho_P = \rho_N = 15$ years^{-1}, while the interaction types and proliferative ability of the E population were varied: (a) $\alpha_{EP} = \alpha_{PE} = 0$ (neutralism) and $\rho_E = 15$ years^{-1}; (b) $\alpha_{EP} = 10$, $\alpha_{PE} = 2$ (cooperation) and $\rho_E = 15$ years^{-1}; (c) $\alpha_{EP} = -5$, $\alpha_{PE} = -2$ (competition) and $\rho_E = 15$ years^{-1}; (d) $\alpha_{EP} = \alpha_{PE} = 0$ and $\rho_E = 16$ years^{-1} (population E has a proliferative advantage over P and N cells); (e) $\alpha_{EP} = 10$, $\alpha_{PE} = 2$ and $\rho_E = 16$ years^{-1}; (f) $\alpha_{EP} = -5$, $\alpha_{PE} = -2$ and $\rho_E = 16$ years^{-1}. The initial conditions are given by Eq. (2.45). Each plot is shown at time $t = 9$ years. Simulations were produced using a Forward-Euler time stepping scheme with a time-step of 0.0005 years and finite differences with a spatial mesh size of 0.25 mm.

2.3.2 Travelling Wave Analysis

Simulations show that the model of interacting GBM sub-populations given by System (2.1)-(2.7) exhibits a variety of travelling wave fronts, the properties of these depending on the model parameters and initial conditions. Figure 2.5 shows some examples of fronts that can emerge from simulations with a variety of parameter values, but the same initial conditions, given by

$$E_0(x) = P_0(x) = N_0(x) = \begin{cases} K/3, & \text{for } x \in [0, 10] \\ 0, & \text{for } x > 10. \end{cases} \tag{2.45}$$

In Fig. 2.5a the diffusion coefficients and proliferation rates of each population are the same, that is $D_E = D_P = D_N = D$ and $\rho_E = \rho_P = \rho_N = \rho$, and there are no interactions between the amplified sub-populations, since $\alpha_{EP} = \alpha_{PE} = 0$. As mentioned earlier in this chapter, the system can be reduced to a single equation governing the total population of cells, $T = E + P + N$, in this case, which is the well-known PI model or Fisher-KPP equation. Travelling wave solutions of this model have been widely studied over the years and, thus, we know the speed of the travelling front of the total population of cells is given by $2\sqrt{\rho D}$. Similarly, in Figs 2.5d and 2.5f the type of interactions and increased proliferation rate of population E allow these cells to dominate at the travelling front with the other two populations absent; this problem is effectively described by a single Fisher-KPP equation again, so that this invading front has speed $2\sqrt{\rho_E D_E}$. Validation of these wavefront speeds are shown in Fig. B.1 in Appendix B.

The examples in Fig. 2.5 all exhibit a single travelling front comprising different proportions of each sub-population of cells that depend on the initial conditions and parameter values used. In addition to these types of waves, there are some instances where a second travelling front can emerge behind the leading invading front.

The plots in Fig. 2.6 show an example where two travelling fronts emerge from the initial conditions,

$$E_0(x) = \begin{cases} K/2, & \text{for } x \in [0, 10] \\ 0, & \text{for } x > 10 \end{cases} \tag{2.46}$$

$$P_0(x) = \begin{cases} K/2, & \text{for } x \in [0, 10] \\ K, & \text{for } x \in (10, 20] \\ 0, & \text{for } x > 20 \end{cases} \tag{2.47}$$

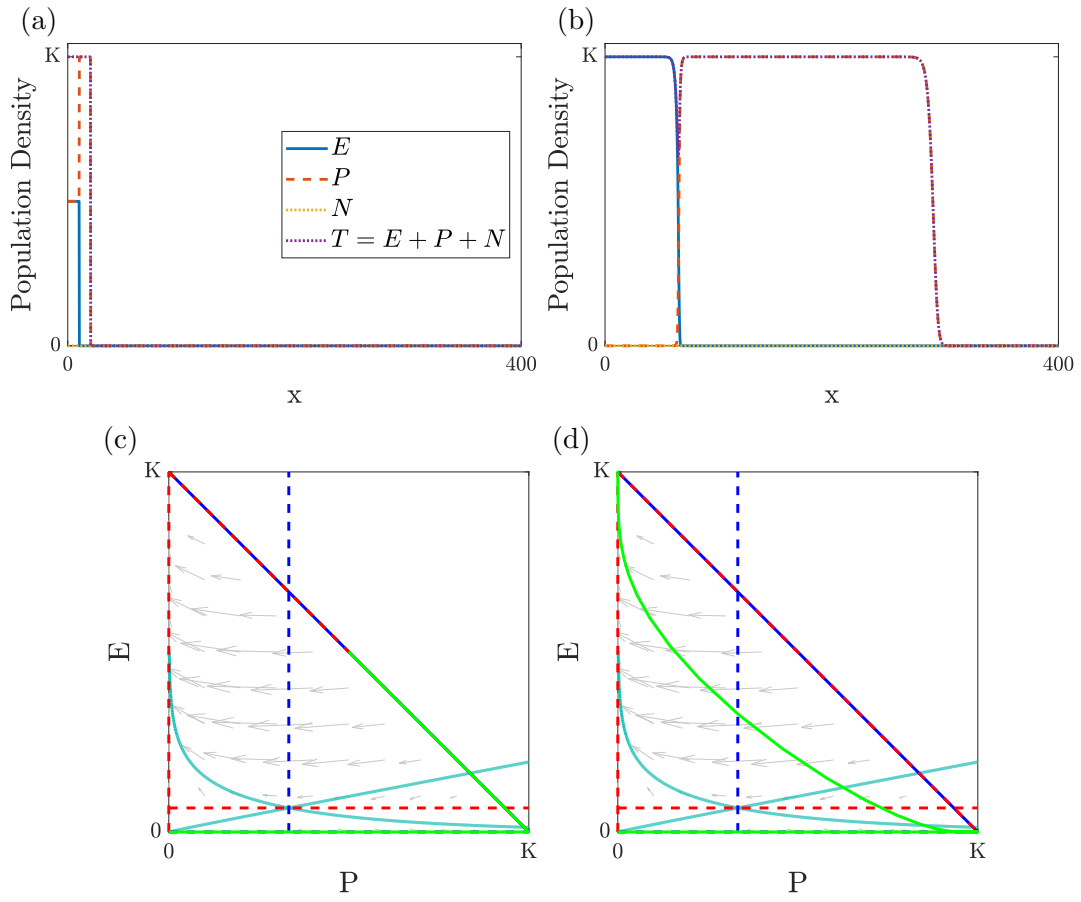


Figure 2.6: Plots showing a wave of E cells invading behind a front of the P population. (a) Plot showing the initial conditions given by Eq.s (2.46)-(2.48). (b) Plot showing the model solution at $t = 8$ years, where we can see a wave of E invading behind the travelling front of P cells. Plots (c) and (d) show E-P phase portraits with separatrices and nullclines shown as previously and the (c) initial conditions and (d) model solution at $t = 8$ years overlaid in green. The parameters used were: $D_E = 25$ and $D_P = 20$ mm^{2}year^{-1}; $\rho_E = 15$ and $\rho_P = 16$ years^{-1}; and $\alpha_{EP} = -15$ and $\alpha_{PE} = -3$. Simulations were produced using a Forward-Euler time stepping scheme with a time-step of 0.0005 years and finite differences with a spatial mesh size of 0.25 mm.

and

$$N_0(x) = 0, \text{ for all } x, \tag{2.48}$$

with one front invading behind the other. In this example, the E population of cells forms a travelling wave that invades space already occupied by the P population, despite the P cells having a higher proliferative ability. This phenomenon occurs due to the competitive interactions between these two populations and the choice of initial conditions. Two separate regions emerge from the initial conditions, one occupied by population E and the other by P, and we see that at the interface of these two regions the total population does not reach K as we might

expect. From Fig. 2.6d, we see that steady states along the line $E + P = K$ are unstable in this case meaning the trajectory connecting the steady states $(E, P) = (K, 0)$ and $(0, K)$ remains below this line and the strong competitive effect of E on population P drives the solution in this region towards the steady state where E dominates, allowing the front of E cells to advance.

The variety of travelling waves arising in simulations of our model and the observation that these emerging waves depend on the initial conditions, could be used to inform the design *in vitro* experiments. For example, different proportions of populations of EGFR and PDGFRA amplified glioblastoma cells could be seeded in invasion assays or cell cultures in various spatial configurations and the patterns of invasion that emerge could then be studied. Such experimental results could provide information about the type of interaction occurring between these amplified sub-populations and also be used to validate our model.

After observing this variety of travelling waves in our model simulations, we are interested in studying them further. For simplicity, we study in 1D and restrict our analysis to the case where only the E and P populations are present and the N population of cells is absent and fix $N \equiv 0$. Introducing the travelling wave coordinate, $z = x - ct$, where c is the wave speed, which we assume to satisfy $c > 0$, and substituting $E(x, t) = E(x - ct) = \bar{E}(z)$, we have

$$\frac{\partial E}{\partial t} = -c\bar{E}' \quad \text{and} \quad \frac{\partial^2 E}{\partial x^2} = \bar{E}'',$$

where $'$ notation denotes differentiation with respect to the travelling wave coordinate z. Repeating this for the P population of cells and dropping the bar notation for simplicity, we find Eq.s (2.1)-(2.2) (with $N = 0$) are transformed to

$$-cE' = D_E \left(1 - \frac{P}{K}\right) E'' + D_E \frac{E}{K} P'' + f_E(E, P), \tag{2.49}$$

$$-cP' = D_P \left(1 - \frac{E}{K}\right) P'' + D_P \frac{P}{K} E'' + f_P(E, P). \tag{2.50}$$

According to the types of travelling waves identified in our simulations, these equations are to be solved according to the following boundary conditions:

- **Case 1:** $E \to 0$, $P \to 0$ as $z \to \infty$, that is the wave propagates into space where no other tumour cells are present, and $E \to E^*$, $P \to P^*$, where $E^* + P^* = K$ and $0 \geq E^*, P^* \leq K$ as $z \to -\infty$, the population densities relax to spatially uniform values that sum to the carrying capacity after the wave has passed.

- **Case 2:** $E \to 0$, $P \to 0$ as $z \to \infty$, that is the wave propagates into space where no other tumour cells are present, and $E \to -K/\alpha_{EP}$, $P \to -K/\alpha_{PE}$ as $z \to -\infty$, the population densities relax to the competition co-existence state after the wave has passed.

- **Case 3:** $E \to 0$, $P \to K$ as $z \to \infty$, that is the wave propagates into space occupied by the P population of cells, and $E \to K$, $P \to 0$ as $z \to -\infty$, population E occupies the space after the wave has passed. We note that the opposite is possible, but we only consider this case.

Rearranging and introducing $E_1 = dE/dz$ and $P_1 = dP/dz$, we have a system of four travelling wave ordinary differential equations (ODEs):

$$\frac{d}{dz}\begin{pmatrix} E \\ E_1 \\ P \\ P_1 \end{pmatrix} = \begin{pmatrix} E_1 \\ g_1(E, E_1, P, P_1) \\ P_1 \\ g_2(E, E_1, P, P_1) \end{pmatrix}, \tag{2.51}$$

where,

$$g_1(E, E_1, P, P_1) = \frac{1}{D_E D_P (K - P - E)}(D_E E(cP_1 + f_P) - D_P(K - E)(cE_1 + f_E)), \tag{2.52}$$

$$g_2(E, E_1, P, P_1) = \frac{1}{D_E D_P (K - P - E)}(D_P P(cE_1 + f_E) - D_E(K - P)(cP_1 + f_P)). \tag{2.53}$$

First we look at case 2, which can be treated in the usual way (as detailed in [71]) by linearising about the point $(E, E_1, P, P_1) = (0, 0, 0, 0)$, that is, the steady state $(E, P) = (0, 0)$ and determining the eigenvalues. These are the roots of

$$\begin{vmatrix} -\frac{c}{D_E} - \lambda & -\frac{\rho_E}{D_E} & 0 & 0 \\ 1 & -\lambda & 0 & 0 \\ 0 & 0 & -\frac{c}{D_P} - \lambda & -\frac{\rho_P}{D_P} \\ 0 & 0 & 1 & -\lambda \end{vmatrix} = 0,$$

which are,

$$\lambda_{1,2} = \frac{-c \pm \sqrt{c^2 - 4\rho_E D_E}}{2D_E} \quad \text{and} \quad \lambda_{3,4} = \frac{-c \pm \sqrt{c^2 - 4\rho_P D_P}}{2D_P}.$$

These eigenvalues all have negative real part, since $c > 0$ by assumption. As the

populations must remain non-negative, we require that

$$c^2 - 4\rho_E D_E \geq 0 \quad \text{and} \quad c^2 - 4\rho_P D_P \geq 0$$

to ensure a spiral approach around the point $(0, 0, 0, 0)$ does not occur. Thus, the only possibility for a travelling wave to exist with non-negative E and P is if

$$c \geq \max\left\{2\sqrt{\rho_E D_E}, 2\sqrt{\rho_P D_P}\right\}. \tag{2.54}$$

Following a similar approach, linearising about the point $(E, E_1, P, P_1) = (-K/\alpha_{EP}, 0, -K/\alpha_{PE}, 0)$, we find that the eigenvalues, λ, are the roots of the equation,

$$p(\lambda) = -\lambda^4 + A\lambda^3 + B\lambda^2 + C\lambda + D, \tag{2.55}$$

where,

$$A = \frac{c(D_P\alpha_{PE}(1+\alpha_{EP}) + D_E\alpha_{EP}(1+\alpha_{PE}))}{D_E D_P(\alpha_{EP}\alpha_{PE} + \alpha_{EP} + \alpha_{PE})}, \tag{2.56}$$

$$B = \frac{-c^2\alpha_{EP}\alpha_{PE}}{D_E D_P(\alpha_{EP}\alpha_{PE} + \alpha_{EP} + \alpha_{PE})} + \frac{\rho_E}{D_E\alpha_{EP}} + \frac{\rho_P}{D_P\alpha_{PE}}, \tag{2.57}$$

$$C = \frac{c(1+\alpha_{EP})(1+\alpha_{PE})}{\alpha_{EP}\alpha_{PE} + \alpha_{EP} + \alpha_{PE}}\left(\frac{\rho_E}{D_E^2\alpha_{EP}} + \frac{\rho_P}{D_P^2\alpha_{PE}}\right), \tag{2.58}$$

$$D = \frac{\rho_E\rho_P(1 + (1+\alpha_{EP})(1+\alpha_{PE}))}{D_E D_P\alpha_{EP}\alpha_{PE}}. \tag{2.59}$$

This equation is difficult (or may not be possible) to solve analytically and find an explicit expression for the roots in terms of the model parameters. However, we note that $p(0) > 0$ since α_{EP} and α_{PE} are both negative and satisfy $1 + 1/\alpha_{EP} + 1/\alpha_{PE} > 0$ when the competition co-existence state exists and the other parameters must all be positive. Thus, as $p(\lambda) \to -\infty$ as $\lambda \to \pm\infty$, at least one eigenvalue must be real and negative, while at least one must be real and positive. Therefore, the point $(E, E_1, P, P_1) = (-K/\alpha_{EP}, 0, -K/\alpha_{PE}, 0)$ will be a saddle-like node, depending on the other eigenvalues. Through plotting the graph of $p(\lambda)$ for various parameter values, we deduce that the other two roots of Eq. (2.55) are either both real and negative or complex. We find that there exists a critical wavespeed c^*, where values of $c > c^*$ mean that both eigenvalues are real, negative and distinct, values of $c < c^*$ mean they are complex and if $c = c^*$ then they are real and equal to one another. Fig.2.55 shows plots of $p(\lambda)$ for a set of parameters and three different wavespeeds, where we observe the roots varying as c is varied. Thus, the wavespeed must satisfy $c > c^*$ and condition

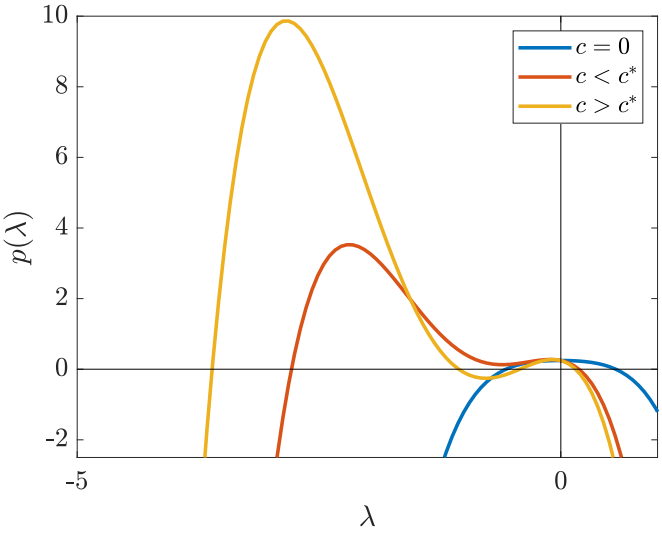


Figure 2.7: Plots of Eq. 2.55 with parameter values $\rho_E = \rho_P = 10$ years^{-1}, $D_E = D_P = 15$ mm^{2}years^{-1} and $\alpha_{EP} = \alpha_{PE} = -3$. The values of c used were 0 (blue), 14.495 (red) and 18.495 (yellow).

(2.54) in order for a trajectory connecting the points $(E, E_1, P, P_1) = (0, 0, 0, 0)$ and $(-K/\alpha_{EP}, 0, -K/\alpha_{PE}, 0)$ to exist.

We note that System (2.51) has a singularity whenever $E + P = K$, thus solving the system with boundary conditions given in cases 1 and 3 requires more care. The singularity can be removed by following a coordinate transformation similarly to [94, 95]. We introduce the coordinate transformation $\xi = \xi(z)$ such that

$$\xi = \int \frac{1}{K - P(z) - E(z)} \mathrm{dz}. \tag{2.60}$$

Defining $E(z) \equiv E(\xi(z))$, $E_1(z) \equiv E(\xi(z))$, $P(z) \equiv P(\xi(z))$ and $P_1(z) \equiv P_1(\xi(z))$, we find that

$$\frac{dE}{d\xi} = (K - P - E)\frac{dE}{dz}, \tag{2.61}$$

along with analogous relations for the other variables. Thus, the original System (2.51) can be re-written as the following non-singular system

$$\frac{d}{d\xi}\begin{pmatrix} E \\ E_1 \\ P \\ P_1 \end{pmatrix} = \begin{pmatrix} (K - E - P)E_1 \\ \frac{1}{D_E D_P}\left(D_E E(cP_1 + f_P) - D_P(K - E)(cE_1 + f_E)\right) \\ (K - E - P)P_1 \\ \frac{1}{D_E D_P}\left(D_P P(cE_1 + f_E) - D_E(K - P)(cP_1 + f_P)\right) \end{pmatrix}, \tag{2.62}$$

together with the conditions

$$0 \leq E(\xi),\ P(\xi) \leq K\ \forall\ \xi \in (-\infty, \infty). \tag{2.63}$$

First we observe that this transformed system has the following steady states; the steady states from the original system,

$$\begin{aligned}(E, E_1, P, P_1) =&(0,0,0,0),\ (0,0,K,0),\ (K,0,0,0),\\ &\left(-\frac{K}{\alpha_{EP}}, 0, -\frac{K}{\alpha_{PE}}, 0\right) \text{ for } -\frac{1}{\alpha_{EP}} - \frac{1}{\alpha_{PE}} \leq 1,\\ &(E^*, 0, K - E^*, 0) \text{ for all } E^* \in (0, K);\end{aligned} \tag{2.64}$$

as well as some additional steady states,

$$\begin{aligned}(E, E_1, P, P_1) =&(K, E_1^*, 0, 0),\ (0,0,K,P_1^*),\\ &\left(E^*, E_1^*, K - E^*, \frac{D_P(K-E^*)}{D_E E^*} E_1^*\right) \text{ for all } E^* \in (0, K).\end{aligned} \tag{2.65}$$

To determine the stability of these steady states, we find the Jacobian matrix of System (2.62), which is given by $J(E, E_1, P, P_1) =$

$$\begin{pmatrix} -\frac{E_1}{K} & 1 - \frac{E+P}{K} & -\frac{E_1}{K} & 0 \\ \frac{\partial}{\partial E}\left(\frac{dE_1}{d\xi}\right) & -\frac{c}{D_E}\left(1 - \frac{E}{K}\right) & \frac{\partial}{\partial P}\left(\frac{dE_1}{d\xi}\right) & \frac{cE}{D_P K} \\ -\frac{P_1}{K} & 0 & -\frac{P_1}{K} & 1 - \frac{E+P}{K} \\ \frac{\partial}{\partial E}\left(\frac{dP_1}{d\xi}\right) & \frac{cP}{D_E K} & \frac{\partial}{\partial P}\left(\frac{dP_1}{d\xi}\right) & -\frac{c}{D_P}\left(1 - \frac{P}{K}\right) \end{pmatrix}, \tag{2.66}$$

where

$$\begin{aligned}\frac{\partial}{\partial E}\left(\frac{dE_1}{d\xi}\right) = \frac{1}{K D_E D_P}&\left(D_E\left(cP_1 + f_P + E\frac{\partial f_P}{\partial E}\right)\right. \\ &\left. + D_P\left(cE_1 + f_E - (K-E)\frac{\partial f_E}{\partial E}\right)\right),\end{aligned} \tag{2.67}$$

$$\frac{\partial}{\partial P}\left(\frac{dE_1}{d\xi}\right) = \frac{1}{K D_E D_P}\left(D_E E\frac{\partial f_P}{\partial P} - D_P(K-E)\frac{\partial f_E}{\partial P}\right), \tag{2.68}$$

$$\frac{\partial}{\partial E}\left(\frac{dP_1}{d\xi}\right) = \frac{1}{K D_E D_P}\left(D_P P\frac{\partial f_E}{\partial E} - D_E(K-P)\frac{\partial f_P}{\partial E}\right), \tag{2.69}$$

$$\begin{aligned}\frac{\partial}{\partial P}\left(\frac{dP_1}{d\xi}\right) = \frac{1}{K D_E D_P}&\left(D_P\left(cE_1 + f_E + P\frac{\partial f_E}{\partial P}\right)\right. \\ &\left. + D_P\left(cP_1 + f_P - (K-E)\frac{\partial f_P}{\partial P}\right)\right).\end{aligned} \tag{2.70}$$

We can now explore the existence of travelling waves in cases 1 and 3 in turn.

Case 1: At $(E, E_1, P, P_1) = (0,0,0,0)$ the eigenvalues of the Jacobian matrix

are

$$\lambda_{1,2} = \frac{-c \pm \sqrt{c^2 - 4\rho_E D_E}}{2D_E} \quad \text{and} \quad \lambda_{3,4} = \frac{-c \pm \sqrt{c^2 - 4\rho_P D_P}}{2D_P},$$

so the wavespeed must satisfy the condition

$$c \geq \max\left\{2\sqrt{\rho_E D_E}, 2\sqrt{\rho_P D_P}\right\}, \tag{2.71}$$

for a travelling wave to exist with E and P remaining non-negative, as found previously. At $(E, E_1, P, P_1) = (E^*, 0, P^*, 0)$, where $E^* + P^* = K$ and $E^*, P^* > 0$, three eigenvalues are zero and the fourth is negative and is given by

$$\lambda = -c\left[\frac{1}{D_E}\left(1 - \frac{E^*}{K}\right) + \frac{1}{D_P}\left(1 - \frac{P^*}{K}\right)\right]. \tag{2.72}$$

As three eigenvalues are zero, the steady state $(E^*, 0, P^*, 0)$ is non-hyperbolic and the linear system will not be enough to learn about the local behaviour around this point. Further non-linear analysis would need to be conducted, which we leave as an open problem in this work. Instead, we explore the behaviour of the system using numerical simulations. Figure 2.8 shows the two types of travelling wave fronts that occur in this case. Fig. 2.8a shows a solution of System (2.62) with the condition on c not satisfied, that is $c < \max\left\{2\sqrt{\rho_E D_E}, 2\sqrt{\rho_P D_P}\right\}$. In this example, $E + P = K$ behind the wave and (E, P) approach $(0, 0)$ in an oscillatory manner, highlighted in Fig. 2.8c. Meanwhile, in Fig. 2.8b $c = \max\left\{2\sqrt{\rho_E D_E}, 2\sqrt{\rho_P D_P}\right\}$ and the two steady states are connected without any oscillations and E and P both remain non-negative.

Case 3: At $(E, E_1, P, P_1) = (0, 0, K, 0)$ and $(K, 0, 0, 0)$, three eigenvalues of the Jacobian matrix are zero and the fourth is negative in both instances and given by

$$\lambda = -\frac{c}{D_P} \quad \text{and} \quad -\frac{c}{D_E},$$

respectively.

Figure 2.9 shows the types of travelling wave fronts that occur with these boundary conditions. Fig. 2.9a shows the solution to System (2.62), for a given value of the wave speed c, starting at values of $(E_1, E, , P_1, P)$ close to the steady state $(0, K, 0, 0)$, the values used were estimated from close to the top of the travelling wave in simulations of the model given by Eq.s (2.1)-(2.2) (with $N = 0$) and were:

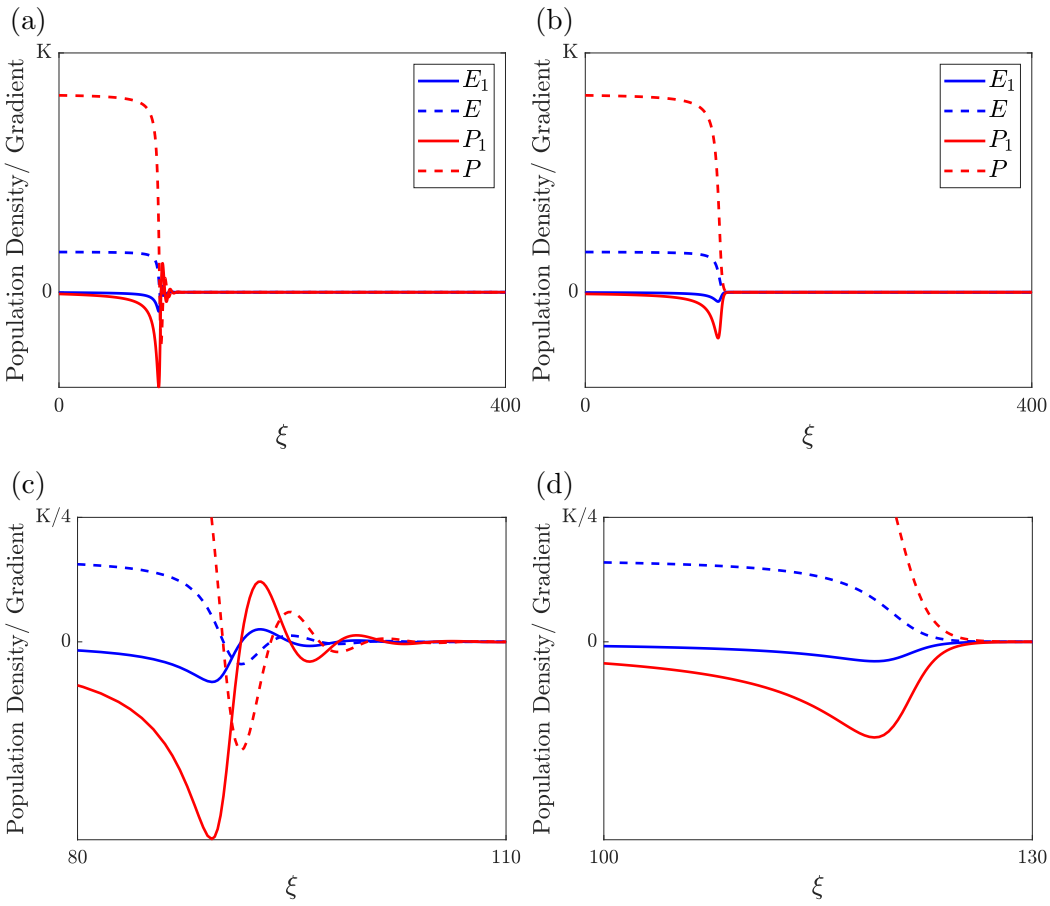


Figure 2.8: Simulations of System (2.62) for two different values of c: (a) and (c) $c = 10$; and (b) and (d) $c = 30$. Each of (a) and (b) shows plots of E_1, E, P_1 and P over $\xi \in [0, 400]$, whereas (c) and (d) show zoomed in portions of these graphs. The parameters used were $\rho_E = \rho_P = 15$ years^{-1}, $D_E = D_P = 15$ mm^{2}year^{-1}, $\alpha_E P = 10$ and $\alpha_{PE} = 2$. Simulations were producing ODE45 in MatLab with initial conditions estimated from simulations of the model given by Eq.s (2.1)-(2.2) (with $N = 0$) and were: $E_1(0) = -342.8044$, $E(0) = 4.013 \times 10^4$, $P_1(0) = -1.7052 \times 10^3$ and $P(0) = 1.9664 \times 10^5$.

$$E_1(0) = -1.6267 \times 10^3, \tag{2.73}$$

$$E(0) = 2.3674 \times 10^5, \tag{2.74}$$

$$P_1(0) = 8.2597 \times 10^{-4}, \tag{2.75}$$

$$P(0) = 0.0015. \tag{2.76}$$

This trajectory then approaches the steady state $(0, 0, 0, K)$ before spiralling towards $(0, 0, 0, 0)$. Increasing c we find a critical value of the wave speed, where a trajectory connects the steady states $(E_1, E, P_1, P) = (0, K, 0, 0)$ and $(0, 0, 0, K)$ in Fig. 2.9b. This type of trajectory corresponds to a travelling wave solution

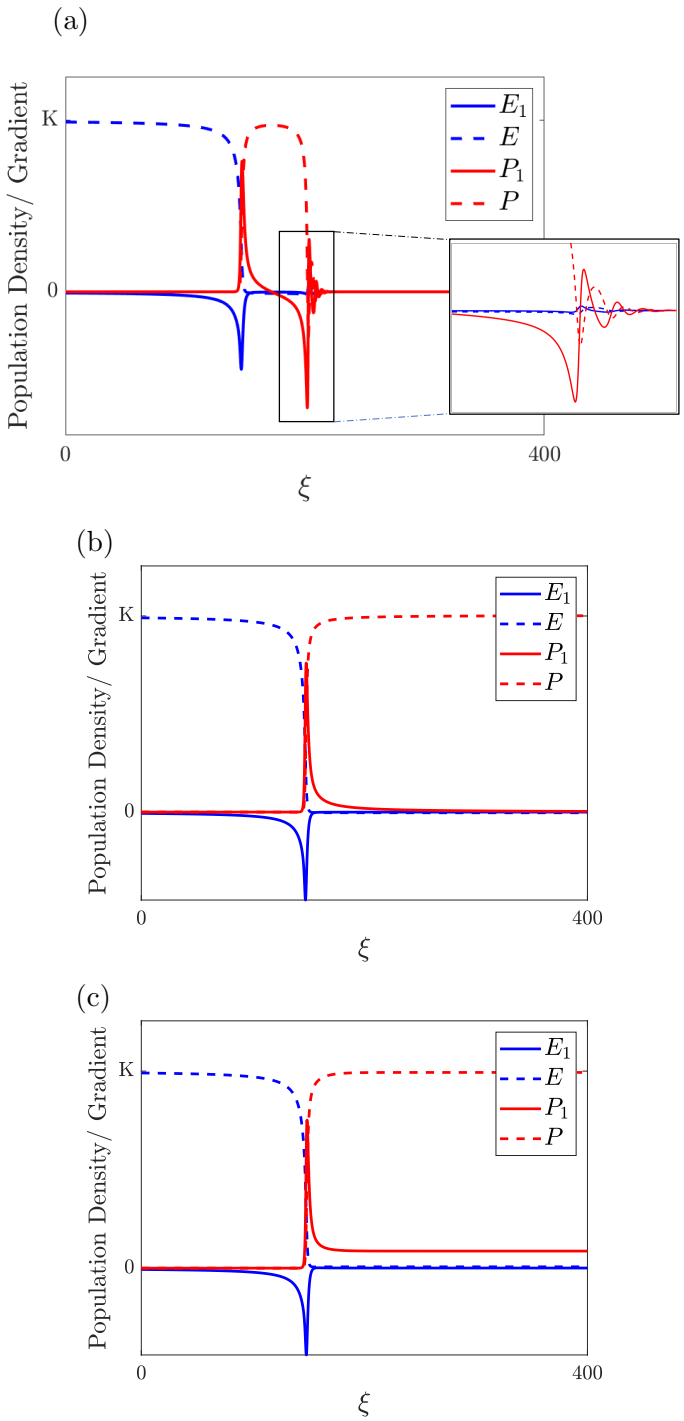


Figure 2.9: Simulations of system (2.62) for three different values of c: (a) $c = 9.738$, (b) $c = 9.838$ and (c) $c = 9.938$. Each shows plots of E_1, E, P_1 and P over $\xi \in [0, 400]$, with (a) also showing a zoomed in portion of the graph. The parameters used were $\rho_E = \rho_P = 15$ years^{-1}, $D_E = D_P = 15$ mm^{2}year^{-1}, $\alpha_E P = -20$ and $\alpha_{PE} = -2$. Simulations were produced using ODE45 in MatLab with initial conditions (2.73)-(2.76).

of the type shown through numerical simulation of the PDE model in Fig. 2.6. Increasing the wave speed above this critical value, we find a family of sharp-

fronted travelling waves connecting $(E_1, E, P_1, P) = (0, K, 0, 0)$ to $(E_1^*, 0, P_1^*, K)$, where E_1^* and P_1^* are non-zero; an example of such a solution is shown in Fig. 2.9c.

2.4 Summary

In this chapter, we have presented a novel mathematical model of the co-evolution of three distinct tumour cell sub-populations to investigate the nature of interactions between cells with two common mutations occurring in GBMs, namely amplification of the genes encoding the EGFR and PDGFRA proteins. We have used a PDE-based formalism, which reduces to the well known PI model [115] - [121] if we assume that these genetic differences do not change the phenotype of the cell populations and instead compose a single phenotypically homogeneous population of cancerous cells. Our model describes the movement of the sub-populations through cross-diffusion terms, where the diffusion of one population is affected by the presence of the other two populations. In Section 2.2.1, we presented a derivation of these movement terms and provided a generalisation to include M populations of distinct cell sub-populations. We also briefly discussed the incorporation of different forms of cross-diffusion into other examples of models comprising two distinct cell populations and the differences to the approach taken in our work.

In our model, the growth terms, f_E, f_P and f_N, are given by Eq.s (2.5)-(2.7). The particular forms of these terms were chosen to allow us to explore the effects of various interactions between different cell types on the growth of GBMs comprised of three distinct sub-populations of cells. We incorporated one factor, namely the joint logistic growth factor, that models the competition for space between all three cell sub-populations and a second factor that models additional interactions between the EGFR and PDGFRA amplified sub-populations. We separated the interactions in this way to highlight that they are distinct and to allow us to clearly explore the impact of a range of types of the additional interactions between the EGFR and PDGFRA amplified sub-populations, while assuming that the cells continue to compete for space in the same way. An alternative approach would be to incorporate our interaction parameters, α_{EP} and α_{PE}, into the joint logistic growth term instead and remove the $(1+\alpha_{PE}P/K)$ and $(1+\alpha_{EP}E/K)$ factors from Eq.s (2.5)-(2.6). For example, candidate growth

terms could be given by

$$f_E(E,P,N) = \rho_E E\left(1 - \frac{E + \alpha_{PE}P + N}{K}\right), \tag{2.77}$$

$$f_P(E,P,N) = \rho_P P\left(1 - \frac{\alpha_{EP}E + P + N}{K}\right), \tag{2.78}$$

with the equation for f_N remaining unchanged, given by Eq. (2.7). Modelling the interactions in this way would lead to potentially quite different model dynamics, as the system would exhibit some different spatially homogeneous steady states (depending on the model parameters) and it would be interesting to explore the model behaviour with these growth terms.

Following on from this, in Section 2.3.1, we studied the dynamics of the spatially homogeneous model system through conducting a phase plane analysis. We began by exploring the dynamics when only the two amplified tumour cell sub-populations were present and explored the effect of various interaction types on the trajectories observed. We examined the competition case in more detail, where we were able to find the equation of the separatrix dividing phase space when the competition co-existence state exists. In Fig. 2.3, we explored how varying the model parameters affects the shape of the separatrix and results in different trajectories of the model with various initial conditions. In particular, we found that a proliferative advantage is not always advantageous for a tumour cell population and there are scenarios where the less aggressive tumour cell population will become the dominant one. We then proceeded to study the spatially homogenous system when all three tumour cell types are present and found that the behaviour is similar to when the non-amplified tumour cells are absent; this is because the more interesting dynamics in our model arise through interactions between the two amplified cell sub-populations.

In Section 2.3.2, we then explored the travelling waves exhibited in simulations of our model, where we found a variety of different waves, some examples of these are shown in Fig.s 2.5 and 2.6. We also conducted a travelling wave analysis for the model when only the two amplified sub-populations are present and found conditions for the various travelling wave solutions to exist. Our results show that the patterns of invasion that occur depend on the model parameters, where we demonstrated that different travelling waves emerge under different types of interactions between the EGFR and PDGFRA amplified sub-populations. Therefore, studying the type of cells present at the invasive edge of glioblastomas could provide an insight into whether these amplified sub-populations are cooperating or competing with each other, for example. Further to this, we also found that the

travelling waves that emerge are dependent on the initial conditions in our model. This means that the type of cells present when a tumour starts growing and the occurrence of any later mutations will result in different patterns of cell invasion, resulting in a variety of amplification patterns observed in GBMs. While this is clearly a complex problem to understand with many factors influencing the invasion of cells and the resulting tumour, the results from our exploration of travelling wave solutions to our model could be used to inform the design of *in vitro* experiments. Such experiments could consist of studying invasion assays or cell cultures of EGFR and PDGFRA amplified cells seeded in different proportions and spatial configurations. Such experiments could provide validation of our model and information about the type of interaction occurring between these amplified GBM sub-populations.

Chapter 3

Preliminary patient data and *in silico* investigations

3.1 Introduction

In this chapter, we use our novel mathematical model of interacting GBM sub-populations to investigate the effects of different interaction assumptions—namely, cooperative, competitive and neutral (no) interactions—on the population level occurrence of EGFR and PDGFRA amplified cells *in silico*.

This chapter begins by describing the process of biopsy collection and analysis for an initial cohort of GBM patients, where we then study population levels found across the data. Next, we introduce terms into the model through which the EGFR and PDGFRA amplified populations arise, as we assume that each tumour comprises only non-amplified tumour cells initially and mutations leading to these sub-populations occur at later times. We then compare the population levels observed across the patient data to model outputs under different interaction assumptions and explore factors affecting the patterns observed in our simulations, such as selection advantages and phylogenetic ordering of mutations, which may also contribute to the levels of EGFR and PDGFRA amplified populations observed in biopsy data. Finally, we conduct a sensitivity analysis of our model and discuss our results and the insight they provide into the evolution of these biologically complex tumours.

The work in this chapter is published in Morris et al. [68].

3.2 A preliminary dataset of image-localised biopsies

Here, details of the image-localised biopsy collection protocol and analysis process are given, as carried out by our collaborators at the Mayo Clinic. We then detail the CNA threshold that we use to determine amplification of a gene in a biopsy in this work and present the data from an initial cohort of GBM patients.

Patients with clinically suspected GBM undergoing preoperative MRI for surgical resection were recruited and the absence of previous treatment was confirmed. Institutional review board approval was obtained, along with written and informed consent from each participant prior to enrollment. During surgery, the surgical team collected an average of 4–5 tissue specimens from each tumour and typically selected targets separated by ≥ 1 cm from different regions of the tumour based on clinical feasibility (e.g., accessibility of the target site, overlying vessels, areas of the brain that directly control function). The location of each biopsy was also recorded by the surgical team to allow for subsequent co-registration with multiparametric magnetic resonance imaging (MRI) datasets. More detail of the biopsy collection protocol can be found in [41].

To determine whether a biopsy sample contains tumour cells with the EGFR and PDGFRA genes amplified, copy number aberration (CNA) values associated with these genes were determined for all tissue samples using array comparative genomic hybridization (aCGH) as described in references [13, 41]. Each tissue sample was then classified as being amplified in a given gene if the corresponding CNA value was greater than a given threshold and not amplified in that gene when below or equal to that threshold. Each biopsy sample, however, is likely to contain a mixture of healthy non-cancerous cells and tumour cells without and with varying degrees of gene amplification. Thus, the CNA value will be based on a mixed signal from a sample containing a mixture of cells with potentially different numbers of copies of the genes of interest and it is unclear what an appropriate threshold should be to determine the gene amplification status, which is a topic widely discussed in the literature [55, 65].

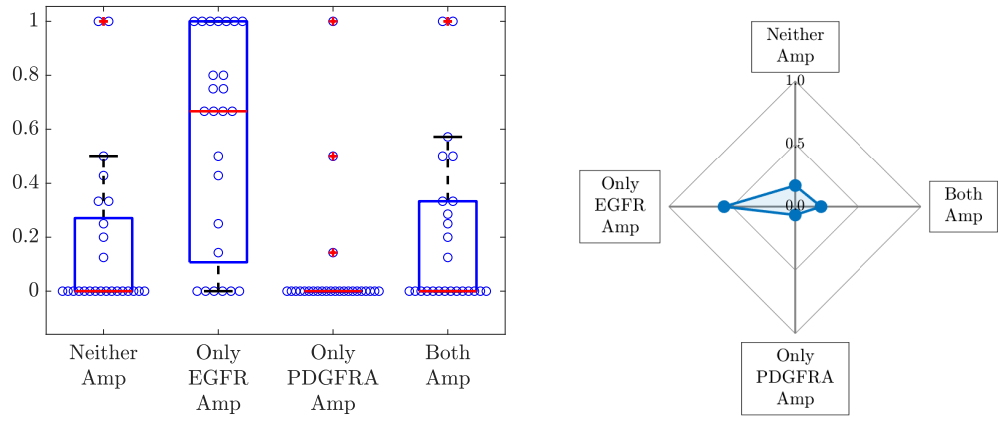
In this work, we choose to use a CNA threshold of 2.2; this threshold is chosen based on some prior knowledge and some assumptions about the levels of EGFR and PDGFRA amplification that we expect to see in our tissue samples. Firstly, diploid cells that are not EGFR or PDGFRA gene amplified will have an associated CNA value equal to 2 [55, 65], which applies to the healthy cells and non-amplified tumour cells in the tissue samples. Secondly, we assume that

the EGFR and PDGFRA amplified cell sub-populations are homogeneous with respect to their gene copy numbers and, therefore, all cells in each of these sub-populations have the same CNA value associated to each of these genes, which we choose to equal 4; this corresponds to each of the alleles in an EGFR amplified cell containing an extra copy of the EGFR gene and similarly for the PDGFRA amplified population. Finally, since the neurosurgeons collect biopsy samples from various regions of each tumour, including the invasive edge where tumour cell density is low, we choose the CNA threshold in a way that will be sensitive to such low densities of EGFR and PDGFRA amplified cell sub-populations; we choose this low density threshold to be 10% of the tissue in a sample and assume that if a biopsy sample consists of 10% or less of either amplified cell sub-population, then the signal will be too low to be detectable in the corresponding CNA value and this sample will not be classed as being amplified in this gene. Therefore, the CNA threshold of 2.2 is derived as follows:

$$\begin{aligned}\text{CNA value of tissue sample} &= \begin{pmatrix}\text{Non-amp fraction of sample} \times \text{CNA value of non-amp cells}\end{pmatrix} \\ &\quad + \begin{pmatrix}\text{Amp fraction of sample} \times \text{CNA value of amp cells}\end{pmatrix} \\ &= (0.9 \times 2) + (0.1 \times 4) \\ &= 2.2\end{aligned}$$

We note that tumour cells can exhibit varying degrees of gene amplification; for example, an EGFR amplified cell sub-population is likely to consist of cells containing a variety of copy numbers of the EGFR gene, with cells containing more than 100 copies in some cases [55]. This means that using a low CNA threshold to determine gene amplification status of a tissue sample in this way could classify a sample containing a very low fraction of "highly amplified" cells as being gene amplified. However, we expect such cases to be rare and choose this CNA threshold to avoid excluding samples that contain cells that are amplified to lower levels.

In this preliminary dataset, a total of 120 biopsies were collected from 25 patients with clinically suspected GBM, with 2–14 collected from each individual. Of these biopsies, 95 samples from 25 patients contained adequate tumour and/or DNA content for EGFR and PDGFRA amplification status to be determined successfully through aCGH analysis. EGFR amplification was the more commonly observed genetic alteration, with 73/95 samples having a CNA value associated with EGFR amplification, whereas 28/95 were determined to be PDGFRA amplified. Of these amplified samples, 22 were found to have amplification of both the



(a) (b)

Figure 3.1: (a) A box plot summarising the proportion of each individual's biopsies that were determined to be amplified in neither gene (Neither), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both of the EGFR and PDGFRA genes (Both Amp) for the 25 patients. Each of the blue circles overlaid on the box plot represents the relevant proportion of an individual's biopsies for each category. The means of these proportions across the 25 patients are shown in (b).

EGFR and PDGFRA genes. For each patient, we then determined the proportion of their biopsies that were found to be amplified in neither gene, only the EGFR gene, only the PDGFRA gene and, finally, both of the EGFR and PDGFRA genes simultaneously. The proportions calculated for each of the 25 patients are summarised as a box plot in Figure 3.1a, where we observe the heterogeneity of amplification patterns observed across the patient cohort. The means of these proportions are shown as a spider plot in Figure 3.1b. Here, we see that the highest mean proportion of biopsies are those amplified in the EGFR gene only, while the means of those amplified in neither and both genes are lower at similar levels. Finally, the mean proportion of biopsies amplified in only the PDGFRA gene is close to zero.

3.3 Introducing EGFR and PDGFRA amplified sub-populations into model simulations

In this work, we assume that each simulated tumour begins as a small population of non-amplified tumour cells (N) and amplification of the EGFR and PDGFRA genes arises via mutations at later times, consistent with a recent phylogenetic study of glioblastomas [108]. Thus, we introduce two new terms, m_E and m_P into the equations for f_E and f_P, through which the sub-populations, E and P,

arise in the model. The modified definitions of the terms f_E and f_P are now,

$$f_E(E,P,N) = \rho_E E\left(1+\alpha_{PE}\frac{P}{K}\right)\left(1-\frac{E+P+N}{K}\right)+m_E, \tag{3.1}$$

$$f_P(E,P,N) = \rho_P P\left(1+\alpha_{EP}\frac{E}{K}\right)\left(1-\frac{E+P+N}{K}\right)+m_P, \tag{3.2}$$

and the term f_N remains unchanged as defined by Eq. (2.7). While these mutation events lead to the creation of a single EGFR or PDGFRA amplified cell and there are likely to be many such events occurring during the growth of a GBM, we assume that each of the EGFR and PDGFRA amplified sub-populations only become established within the tumour at most once. Furthermore, since we are using a PDE model more suited to modelling events on the macroscopic scale rather than single cell events, we account for these successful mutation events by introducing a small population of the mutated cells as a distribution. This means that the mutated cell population actually started growing a small amount of time before being introduced in our model. However, we assume that this will not have affected the other tumour cell populations present and they only begin interacting and competing for space and resources once the population is of a certain size. We, therefore, choose $m_E = m_E(x,t,N,N_E)$ and $m_P = m_P(x,t,N,N_P)$ to be of the following form,

$$m_E(x,t,N,N_E) = \frac{100}{\sqrt{\pi}}\delta(t-t_E^*(N,N_E))e^{-|x-x_E^*|^2}, \tag{3.3}$$

$$m_P(x,t,N,N_P) = \frac{100}{\sqrt{\pi}}\delta(t-t_P^*(N,N_P))e^{-|x-x_P^*|^2}, \tag{3.4}$$

where $\delta(\cdot)$ is the Dirac delta function, $t_E^*(N,N_E)$ and $t_P^*(N,N_P)$ are the times at which populations of EGFR and PDGFRA amplified cells, E and P, are introduced as Gaussian distributions centred at x_E^* and x_P^*, respectively. The introduction times, $t_E^*(N,N_E)$ and $t_P^*(N,N_P)$, are defined as

$$t_E^*(N,N_E) = \inf\left\{t>0\,\middle|\,\int_0^L N(x,t)\,\mathrm{d}x = N_E\right\}, \tag{3.5}$$

$$t_P^*(N,N_P) = \inf\left\{t>0\,\middle|\,\int_0^L N(x,t)\,\mathrm{d}x = N_P\right\}, \tag{3.6}$$

so that each sub-population is introduced when the non-amplified tumour population ($N(x,t)$) has grown to a chosen total size, N_E or N_P (measured in cells/mm^2). Our choices for N_E or N_P are discussed further in Section 3.4.2. Although we have used Gaussian distributions, the specific form of distribution does not sig-

nificantly change our model simulations, as long as they were chosen consistently, due to the smoothing effects of diffusion in the model.

Throughout this work, we continue to consider the model—now defined by Equations (2.1)-(2.3), (2.7) and (3.1)-(3.6)—on a one-dimensional cartesian domain, $x \in [0, L]$, with zero flux boundary conditions at $x = 0$ and L. We use the initial conditions given by,

$$E(x,0) = 0, \quad P(x,0) = 0 \quad \text{and} \quad N(x,0) = \frac{100}{\sqrt{\pi}} e^{-|x - x_N^*|^2}, \tag{3.7}$$

where E, P and N are the concentrations of each of the tumour cell sub-populations (cells/mm^3) and x_N^* defines the centre of the initial distribution of type N tumour cells. Thus, we initiate our simulations with no E or P cells present and a small population of type N cells.

For the remainder of this thesis, we consider a domain length of $L = 200$ mm. Since we are only interested in running simulations to a biologically relevant size (discussed further in Section 3.4.2), and all populations are introduced close to its centre, this domain is sufficiently large that tumour growth remains far from the boundaries, hence avoiding boundary condition artefacts. All simulations are produced using MatLab R2017a to implement a finite difference scheme in space (uniform mesh size of 0.25 mm) and a Forward Euler time step of 1/1500 years.

3.4 Results

3.4.1 Our model predicts that distinct competing sub-populations of tumour cells can coexist in the same tumour region

Intuitively, we expect that cell populations actively competing with one another may be less likely to coexist within the same region of a tumour and that their coexistence may indicate a cooperative relationship, as Snuderl et al. [107] suggest after finding intermingled sub-populations of EGFR and PDGFRA amplified sub-populations in a small number of GBM samples. In our model, however, we find that EGFR and PDGFRA amplified sub-populations can be found to coexist in areas of the tumour region despite actively competing with one another; we find that any $\bar{E}, \bar{P}, \bar{N} \geq 0$ satisfying $\bar{E} + \bar{P} + \bar{N} = K$ is a spatially homogeneous steady state and can be connected to other spatially homogeneous steady states satisfying the same condition, or the trivial steady state $\bar{E} = \bar{P} = \bar{N} = 0$, by

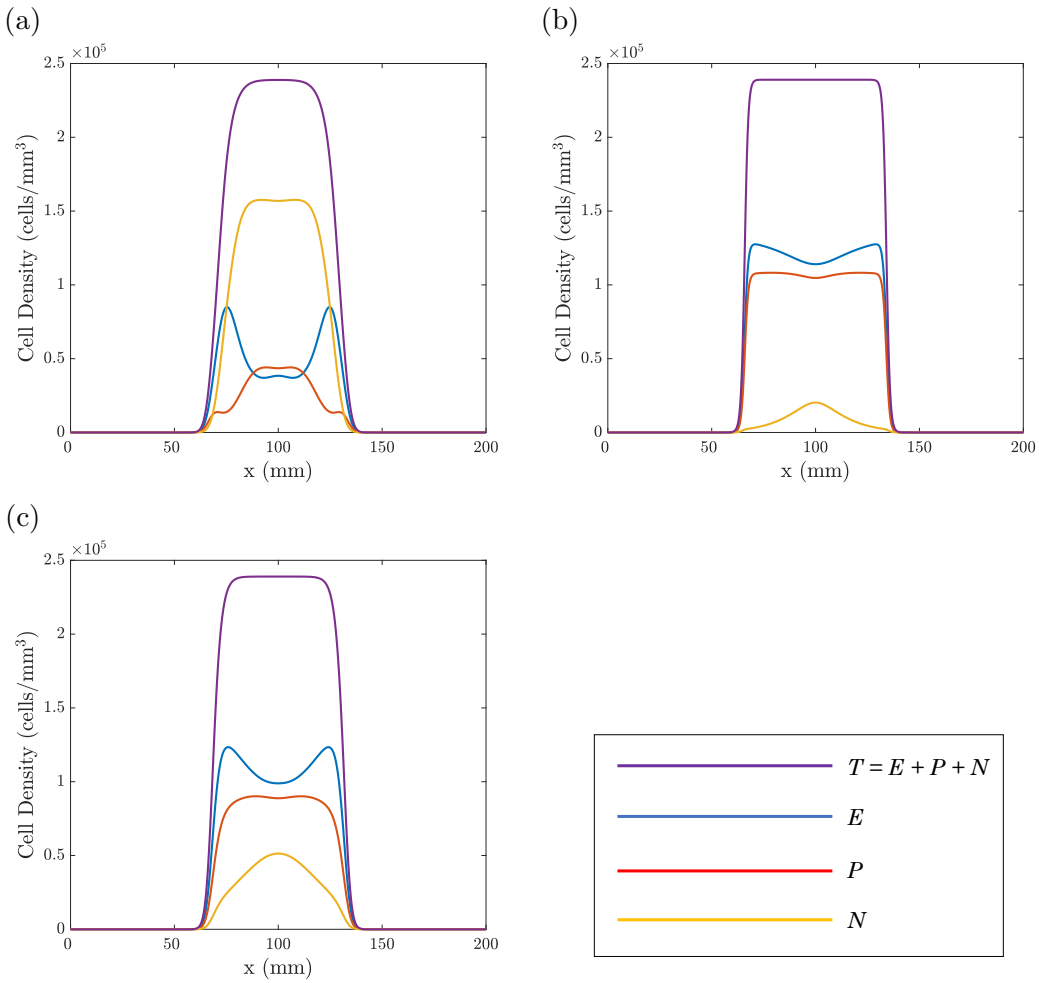


Figure 3.2: Simulations in 1D of the model given by Eq.s (2.1)–(2.3), (2.7) and (3.1)–(3.6) and initial conditions (3.7) using the finite difference scheme described in Section 3.3 with parameters: $\rho_E = 35.4$, $\rho_P = 33$ and $\rho_N = 30$ /year; $D_E = D_P = D_N = 30$ mm^2/year; $K_E^* = K_P^* = K_N^* = K = 2.39 \times 10^5$ cells/mm^3; $x_E^* = x_P^* = x_N^* = 100$ mm; $t_E^* = 0.027$, $t_P^* = 0.001$ years. The interactions in each simulation are chosen to be (a) competition, $\alpha_{EP} = \alpha_{PE} = -5$, (b) cooperation, $\alpha_{EP} = \alpha_{PE} = 5$ and (c) neutralism, $\alpha_{EP} = \alpha_{PE} = 0$. Each simulation is plotted at $t = 0.7$ years.

travelling wave-like solutions expanding outwards from the origin of the tumour. Indeed, such co-occurrence of EGFR and PDGFRA amplified cell sub-populations can be observed with competitive, cooperative or neutral interactions; an example of simulations with different α_{EP} and α_{PE} values to represent each of these interaction types is shown in Figure 3.2, where such co-occurrence is observed.

While this demonstrates that the coexistence of EGFR and PDGFRA amplified cell populations in one known region of a tumour can occur when they are competing, cooperating and evolving neutrally, it highlights the need for more information to determine the nature of such interactions between co-occurring cells *in vivo*. In this chapter, we hope to shed some light on this by studying

amplification patterns observed across image-localized biopsies from the initial cohort of patients and the patterns that emerge in simulations of our model—given by Eq.s (2.1)–(2.3), (2.7) and (3.1)–(3.6) and initial conditions (3.7)—under different interaction assumptions.

3.4.2 Simulation results

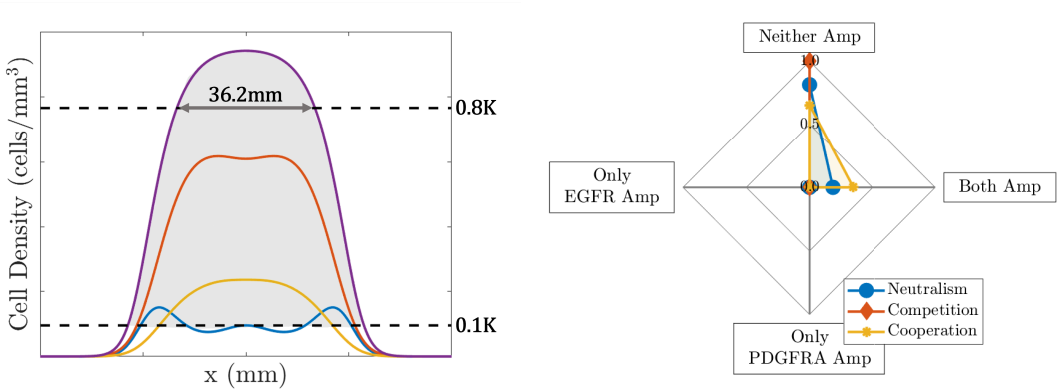
We begin by assuming that amplification of the EGFR or PDGFRA gene does not result in either sub-population, E or P in the model, acquiring any selective advantages over non-amplified cells. In other words, we choose all proliferation and invasion parameters to be the same for each population, i.e., $\rho_E = \rho_P = \rho_N = \rho$ and $D_E = D_P = D_N = D$. Since the biopsy data (Figure 3.1b) is the mean of a cohort containing 25 individual tumours, which will vary in their proliferative and invasive potential (quantified by ρ and ρ/D in the PI model [6]), we use a variety of ρ and D pairs to reflect the heterogeneity seen in the patient cohort. We therefore produce four simulations using two values for each of ρ and D to mirror the range of parameters observed in unpublished patient databases and assume this cohort are similarly distributed; to represent high parameter values we use $\rho = 30$/year and $D = 30\text{mm}^2$/year and for low values we use $\rho = 3$/year and $D = 3\text{mm}^2$/year as in [36]. We also assume that any interactions between the EGFR and PDGFRA amplified cells affect each sub-population to the same degree, i.e., $\alpha_{EP} = \alpha_{PE} = \alpha$. To represent competition and cooperation, we simulate with $\alpha = -5$ and 5, respectively, while for neutralism $\alpha = 0$. For now, we assume that all populations, E, P and N, are introduced at the centre of the domain ($x_E^* = x_P^* = x_N^* = 100\text{mm}$) and that the mutation events for the introduction of EGFR and PDGFRA amplified sub-populations occur at the same time ($t_E^* = t_P^*$, i.e., $N_E = N_P$). Recall that the introduction times depend on the size of the non-amplified tumour population (N), through Eq.s (3.5) and (3.6). Since the time to reach a specific tumour size depends on ρ, rather than introduce these populations at a specified point in time, we introduce them after a given number of proliferation events have occurred, i.e., a specified stage in the evolution of the tumours. In this case we introduce the amplified populations after the tumour has grown to six times its initial size, that is $N_E = N_P = 6N_I$, where $N_I = \int_0^L N(x,0)\,\mathrm{d}x$. While this may seem particularly early in the evolution of our simulated tumours, it is necessary to be able to investigate the amplification patterns we observe under different interactions in our model simulations; if we introduce the mutated populations later, e.g., when the N population has grown to $11N_I$, we do not see them growing to detectable levels in our simulations.

In order to test whether neutral, competitive or cooperative interactions between EGFR and PDGFRA amplified sub-populations best describe the patterns observed in the biopsy data, we run numerical simulations of the model—given by Eq.s (2.1)–(2.3), (2.7) and (3.1)–(3.6) and initial conditions (3.7) using the finite difference scheme described in Section 3.3—to a biologically relevant size and compare the outputs to the mean proportions of biopsies amplified in neither gene, only the EGFR gene, only the PDGFRA gene and both of the EGFR and PDGFRA genes shown in Figure 3.1b.

We choose a biologically relevant size to represent a typical size of a GBM tumour at the time of diagnosis, shortly after which patients will usually undergo surgery to remove as much of the tumour as feasible, with biopsy samples being collected at this time. One way to measure the typical size of a GBM at diagnosis is to segment the tumour volume visible on a T1-weighted MRI with gadolinium contrast (T1Gd MRI), a type of scan that shows the most dense area of the tumour and is typically used in the process of diagnosing a patient with a GBM. From this, the diameter of the volume-equivalent sphere can then be computed and used as a measure to indicate the size of the tumour lesion, with the average diameter at diagnosis being 36.2 mm in unpublished data. Following Swanson et al. [116], we relate the tumour volume visible on a T1Gd MRI to the volume of tumour that has a tumour cell density greater than 80% of the carrying capacity. Therefore, we choose to run simulations until the width of the total tumour cell population, $T = E + P + N$, above the $0.8K$ threshold is 36.2 mm and use this as a proxy for the size of a tumour at the time of diagnosis.

We then run the numerical simulations until they reach this biologically relevant size with the parameter sets described above. Since the patient data is for the proportions of biopsies containing EGFR and PDGFRA amplified cells above a given density threshold, which we chose to be 10% of the tissue sample, we define an equivalent measure for each simulated tumour by integrating solutions across the whole tissue domain. For example, the proportion of the tumour with only EGFR amplified, A_E, is calculated as

$$\begin{aligned} A_E(t) &= \frac{\int_0^L H(E(x,t) - 0.1K)H(0.1K - P(x,t))\,\mathrm{d}x}{\int_0^L H(E(x,t) + P(x,t) + N(x,t) - 0.1K)\,\mathrm{d}x}, \\ &\approx \frac{\text{Number of mesh points with } E > 0.1K \text{ and } P < 0.1K}{\text{Number of mesh points with } T > 0.1K}, \end{aligned} \tag{3.8}$$



(a) (b)

Figure 3.3: (a) Schematic illustrating the biologically relevant size that tumours are simulated to and area (shaded in grey) indicating the points with the total tumour cell population (purple curve) above the threshold of 10% of the carrying capacity. Other curves represent the individual tumour cell populations; E (blue), P (red) and N (yellow). (b) Plot showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified under different interactions when we assume that the sub-populations are dynamically the same, i.e., all populations have the same proliferation and invasion parameters, $\rho = \rho_E = \rho_P = \rho_N$ and $D = D_E = D_P = D_N$, and E and P populations are introduced at the same position and time, $x_E^* = x_P^* = x_N^* = 100$mm and $N_E = N_P = 6N_I$, as described in Section 3.4.2.

where $H(\cdot)$ is the Heaviside step function. The proportions with neither gene, only the PDGFRA gene and both genes amplified, $A_N(t)$, $A_P(t)$ and $A_B(t)$, are defined and calculated in a similar way. We illustrate this schematically in Figure 3.3a.

The mean proportions from the simulations run using each of the four ρ and D pairs with the different cases of competition, cooperation and neutralism are shown in Figure 3.3b. From this plot, we see that all simulations are classed as neither gene amplified in the competitive case and this proportion decreases as we move through the neutralism case to the cooperative case and the proportion with both genes amplified increases, which is as we would intuitively expect to see. We notice that in all three cases the proportions of simulations with only one of the genes amplified are zero; again, this is expected as populations E and P have the same proliferation and invasion parameters (ρ and D) and are both introduced at the same time and place and, thus, are effectively the same so we expect to see them together (or not at all in the competitive case).

Clearly, the patterns of neither, only EGFR, only PDGFRA and both amplified proportions we see in these simulations do not reflect the patterns of am-

plification we see in the biopsy data in Figure 3.1b; one obvious difference is the proportion of only EGFR amplified biopsies, which is above 0.5 in the data and is 0 in the simulations shown in Figure 3.3b. Since EGFR and PDGFRA amplified cells are not exclusively found in the same biopsies in the patient data and the proportions of biopsies with only one gene amplified also differ, this indicates the two populations must differ in their dynamics in some way. There are several possible ways that differences between the EGFR and PDGFRA amplified sub-populations can occur in our model: by giving them a selective advantage (changing ρ_E, ρ_P, D_E or D_P); by changing the phylogenetic ordering of mutations (changing t_E^* or t_P^*); by changing the location that mutations arise in the evolving tumour (changing x_E^* or x_P^*); by changing the strength of interaction felt by each population (changing α_{EP} or α_{PE}); or, finally, by changing any combination of these factors.

Selection advantages

The amplification of EGFR and PDGFRA genes are often considered to be among the key mutations driving oncogenesis and tumour growth [16, 107]. Since these genes are both members of the RTK family of cell surface receptors that play an important role in the regulation of cell proliferation, metabolism and survival [30] and tumours identified to be EGFR amplified have shown to be more invasive [123, 124], it is reasonable to consider that amplification of these genes may drive the growth of tumours through increasing the intrinsic proliferative and invasive abilities of these cell sub-populations. We can explore the effects of this by giving the EGFR and PDGFRA amplified sub-populations in our model different selection advantages through changing the appropriate parameters, namely ρ_E, ρ_P, D_E and D_P.

As we see a much higher proportion of biopsies with only EGFR amplified in the patient data (Figure 3.1b), this could indicate that EGFR amplified cells have a selective advantage over the PDGFRA amplified cells and those with neither gene amplified. Thus, we explore the effects of giving EGFR amplified cells invasive and proliferative advantages; in Figure 3.4, plots are produced from simulation results in the same way as previously described, but with (a) a 50% invasive advantage for the EGFR amplified cells and (b) additionally with a proliferative advantage. In both of these cases, there are now large proportions of simulations with EGFR amplified and without PDGFRA amplified cells present, particularly in the competitive and neutral cases. In both competitive cases in Figures 3.4a and 3.4b, we observe nowhere where both of the EGFR and PDGFRA genes

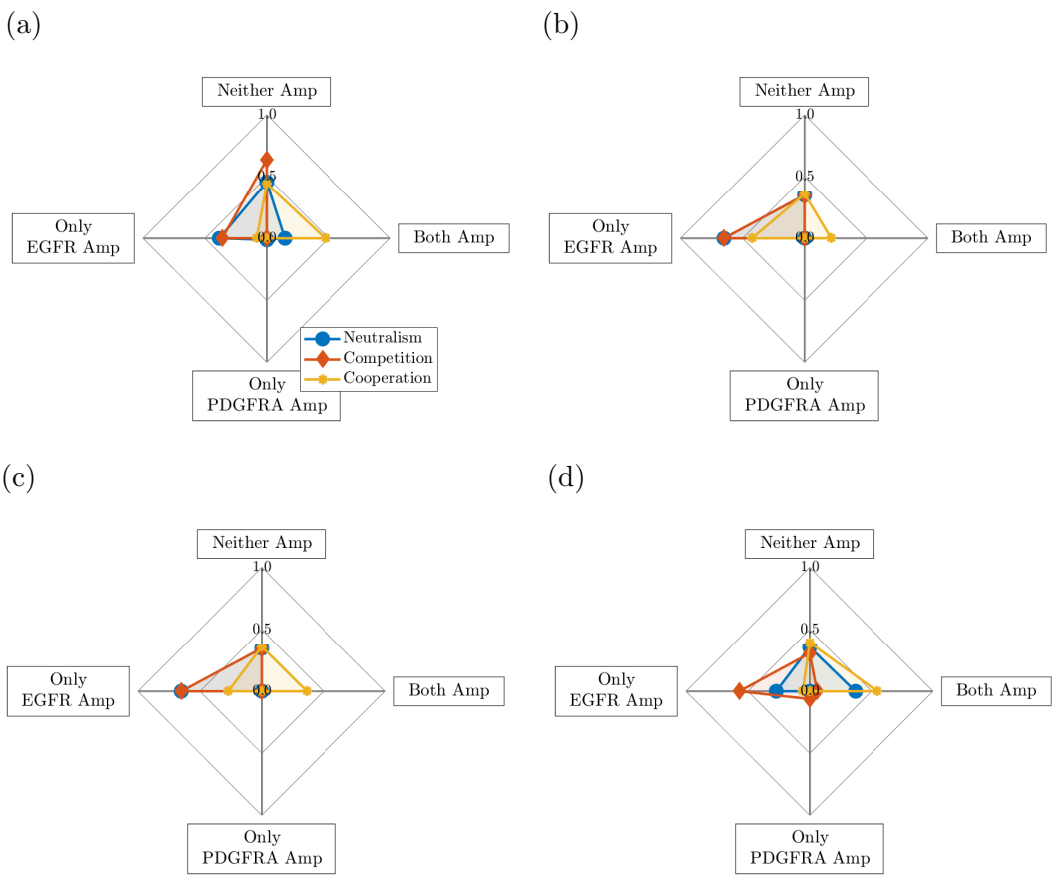


Figure 3.4: Plot showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified under different interactions when the E and P sub-populations are given various selection advantages: (a) EGFR 50% invasive advantage ($D_E = 1.5D_N$, $D_P = D_N$); (b) EGFR 50% proliferative and invasive advantage ($\rho_E = 1.5\rho_N$, $\rho_P = \rho_N$, $D_E = 1.5D_N$ and $D_P = D_N$); (c) EGFR 50% proliferative and invasive advantage, PDGFRA 50% invasive advantage ($\rho_E = 1.5\rho_N$, $\rho_P = \rho_N$, $D_E = 1.5D_N$ and $D_P = 1.5D_N$); (d) EGFR 50% proliferative and invasive advantage, PDGFRA 50% proliferative advantage ($\rho_E = 1.5\rho_N$, $\rho_P = 1.5\rho_N$, $D_E = 1.5D_N$ and $D_P = D_N$)

are found to be amplified, unlike the patient data in Figure 3.1b where this was approximately 20% of biopsies. Meanwhile, the corresponding cooperative cases both gave the highest level of points with both genes amplified and the lowest levels with only EGFR amplified. In Figure 3.4, we also explored the effects of affording the PDGFRA amplified sub-population a (c) 50% invasive and (d) 50% proliferative advantage over non-amplified cells while giving EGFR amplified cells the same advantages as in (b). These resulted in qualitatively similar amplification patterns to those observed in (b) and (a), respectively, with the exception of the competitive case in (d) where a small proportion of simulations with only PDGFRA amplified were observed. Further plots exploring the effects of selec-

tion advantages can be found in the supplementary material in Figure C.1 and are not presented here for brevity. We find that giving either, or both, of the amplified populations invasive and proliferative advantages over non-amplified cells improves the amplification patterns we see in simulations with respect to the biopsy data and, in general, the competition and neutral cases fit the biopsy data better than the cooperative case, which generally results in higher proportions of tumours with both genes amplified. Next we move on to explore the effects that the timing of mutations have on the results we see.

The phylogenetic ordering of mutations

In all simulations presented up until this point, we have assumed that the mutations leading to the establishment of EGFR and PDGFRA amplified cell sub-populations occur at the same time. While this could be the case, a study reconstructing the phylogeny of GBMs identified EGFR and PDGFRA amplification as early and late events during tumour progression [108]. From an analysis of multiple spatially distinct samples from 11 GBMs, it was inferred that alterations that were more common occurred earlier in the evolution of the tumour compared to those only present in a smaller subset of cells. Alterations in gene copy numbers on the chromosomes where the EGFR and PDGFRA genes are found tended to occur in the early and middle phases of tumour growth, respectively [108]. Therefore, this timing, or *phylogenetic ordering*, of mutations may be affecting the proportions of EGFR and PDGFRA amplified biopsies across the patient cohort (Figure 3.1b) and so we undertake a brief exploration of the effect it has in our simulations.

Therefore, we assume once more that all cell sub-populations have the same proliferation and invasion parameters, i.e., $\rho = \rho_E = \rho_P = \rho_N$ and $D = D_E = D_P = D_N$. As described before, we then simulate using four different ρ and D pairs and the same assumptions described in Section 3.4.2, with the exception of changing the times that the EGFR (E) and PDGFRA (P) amplified populations are introduced, i.e., t_E^* and t_P^*. In the previous simulations, both populations were introduced after the non-amplified population (N) had grown to six times its initial size, $6N_I$, and so we now choose to investigate the patterns of gene amplification we see when the E and P populations are introduced earlier and later than this. Thus, we define a vector of possible introduction times, $\mathbf{t}^* = (t_1^*, ..., t_7^*)$ as follows: t_i^* is defined as the first time point in our simulations after the growing N population has reached a size of $(i{+}2)N_I$, for $i = 1, ..., 7$. Implementing our model with t_E^* and t_P^* taking each of these values, we gain some insight into

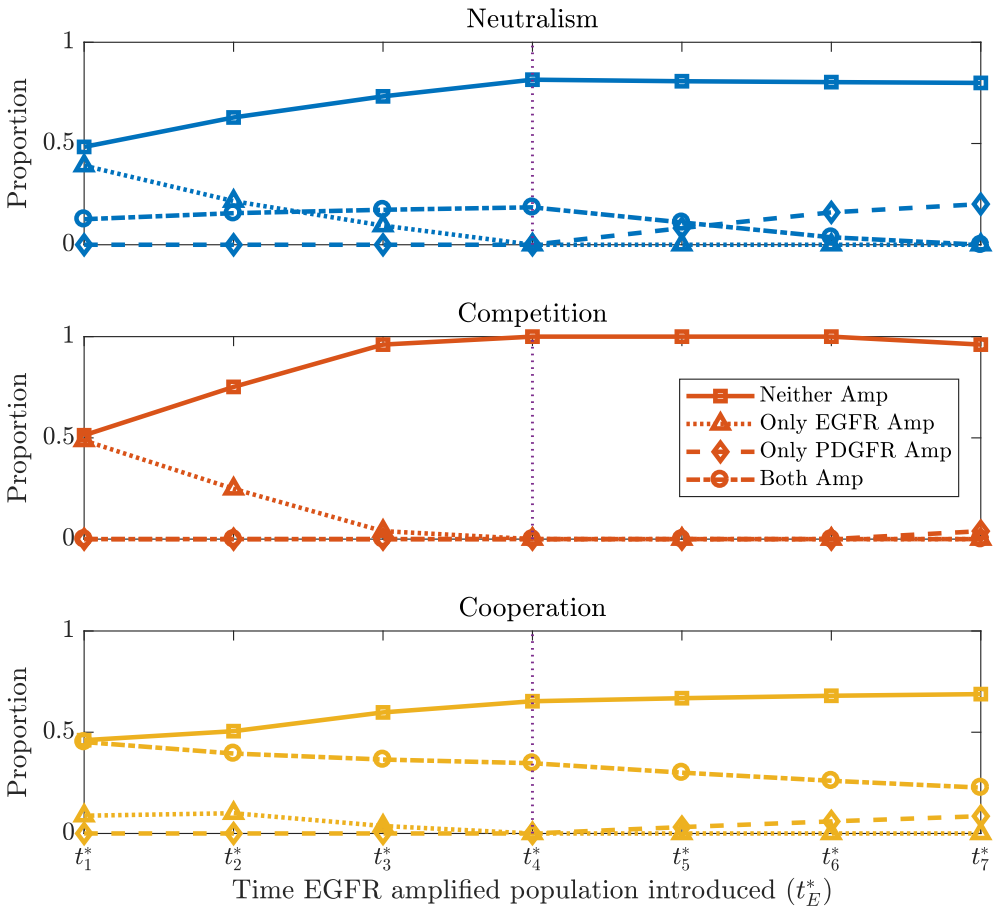


Figure 3.5: Plots showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified as t_E^* is varied, under neutral, competitive and cooperative interactions. The PDGFRA amplified population is introduced at the fixed time $t_P^* = t_4^*$ (denoted by the vertical dotted line) and other parameters and assumptions are as described in Section 3.4.2.

the effect that changing the time of mutations has on our simulations.

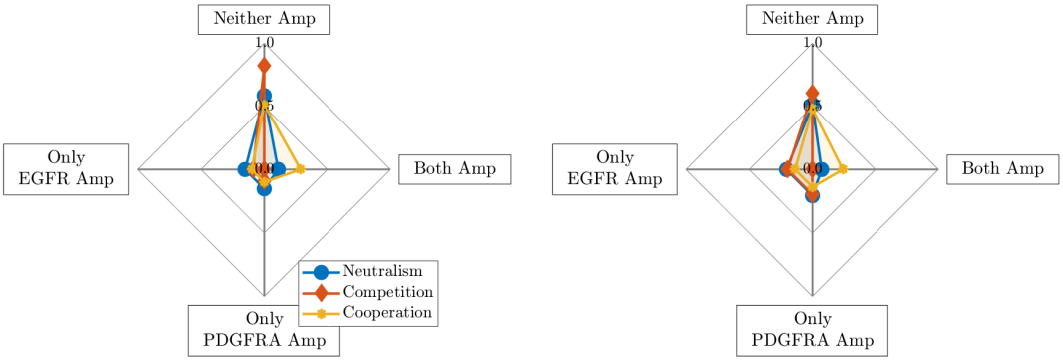
Figure 3.5 shows the average amplified proportions in seven sets of simulations under neutral, competitive and cooperative interaction assumptions, where cells of type P are introduced at $t_P^* = t_4^*$ in each set, but the E population is introduced at each of the seven possible times given by the vector $\mathbf{t}^*$. The points on the graph where $t_E^* = t_4^*$ (marked by the dotted vertical line) are the same data represented in Figure 3.3b as the EGFR and PDGFRA amplified populations are introduced at the same time. From this graph, we see that the proportion with neither gene amplified decreases and that with EGFR amplified increases as the EGFR amplified population is introduced earlier. Meanwhile, as E cells are introduced later than P cells, we see the proportions changing as we would expect, with the only PDGFRA amplified proportion increasing. We also note the slight decrease in the proportion of neither amplified cells when the EGFR amplified population

is introduced at $t_E^* = t_7^*$ in the competitive case, this is because the PDGFRA population is allowed enough time to proliferate to a size where the introduction and competitive interactions of EGFR amplified cells has a smaller relative effect on their proliferation. Again, for brevity we do not present all the simulation results here, further plots can be found in the supplementary material (Figures C.2 and C.3). As with the previous section where we looked at the effect of giving the amplified sub-populations selection advantages, we find that changing the timing of introducing the mutated populations in our simulations does not fit the biopsy data perfectly, but has some of the desired effects. For example, to get proportions more similar to the biopsy data, we need the proportion with neither gene amplified to decrease and the EGFR amplified proportion to increase, which is consistent with earlier EGFR introduction times in the competitive and neutral scenarios, whereas the cooperative case produces a much higher proportion of both amplified.

The location of mutations

Another factor that could result in some biopsies having only EGFR and others only PDGFRA amplified is that the mutations occurred in different places, resulting in the populations occupying different, spatially separated regions of the tumour. In all previous simulations presented in this paper, we assumed that the mutation events leading to the establishment of EGFR and PDGFRA amplified sup-populations in our model occurred in the centre of the growing tumour; this was a reasonable place from which to explore the effects of selection advantages and timing of mutations, since it is where most proliferation is taking place in our model in the early phases of tumour growth and, therefore, where we may expect more mutations to appear. However, cells at the centre of the tumour also experience more competition for space, a growth limiting resource in our model, as this is where the highest tumour cell density is in these early growth phases. Therefore, we now explore the effects of introducing the EGFR and PDGFRA amplified sub-populations away from the centre of the tumour where there is less competition for space.

We use the same parameters and assumptions described in Section 3.4.2, but define a vector of possible introduction locations. In these simulations, the mutated populations are introduced when the growing tumour is small (the N population is only six times its initial size), with the bulk of the tumour being contained within a width of 3mm for each (ρ, D) pair at this time. Thus we choose introduction locations no further than 1.5mm from the tumour centre as



(a) (b)

Figure 3.6: Sub-population distributions when mutations are introduced on opposite sides of the tumour: plots showing mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified when EGFR and PDGFRA amplified sub-populations are introduced (a) 0.5mm to left and right of the center (i.e., $x_E^* = x_3^*$ and $x_P^* = x_5^*$) and (b) 1mm to left and right of the center (i.e., $x_E^* = x_2^*$ and $x_P^* = x_6^*$), respectively.

mutations are unlikely to occur beyond this point where concentrations of tumour cells are very low and, consequently, few proliferation events are occurring. We define a vector of possible introduction locations, $x^* = (x_1^*, ..., x_7^*)$ as follows: $x_i^* = x_c^* + 0.5(i-4)$mm, where x_c^* is the mesh point at the centre of the tumour. We produce simulations with x_E^* and x_P^* each taking the values given by the vector x^*. We show two simulation results in Figure 3.6 showing the proportions when EGFR and PDGFRA amplified populations are introduced (a) 0.5mm and (b) 1mm to the left and right of the tumour center, respectively. In both of these cases we see distinct regions with EGFR and PDGFRA amplification forming in the simulated tumours of equal proportion. We refer the reader to Figures C.4 and C.5 in the supplementary material for further results where we find that introducing each population closer to the tumour centre has a small effect on the amplified tumour proportion and that introducing PDGFRA further away from the EGFR amplified population has similar effects to those presented here in Figure 3.6. As found in the previous two sections, only changing where the amplified sub-populations are introduced does not fully explain the patterns of amplification we observe in the biopsy data, however it does produce some of the desired effects, such as increasing the proportion of simulations with only one of each gene amplified.

3.4.3 LHS-PRCC Sensitivity Analysis

To consider the implications of our model in a more comprehensive manner, in this section we conduct a sensitivity analysis through Latin Hyper-cube Sampling (LHS) and Partial Rank Correlation Coefficients (PRCC). This enables a better understanding of the effects of parameter values on the amplification patterns we observe across our model simulations. In our analysis we include 12 model parameters and assign a uniform probability density function (pdf) to each, with minimum and maximum values chosen in line with the ranges explored in previous sections of this paper; the details of which are outlined below and summarised in Table 3.1.

In Section 3.4.2 we explored the effects of affording the EGFR and PDGFRA amplified sub-populations various selection advantages and changing the phylogenetic ordering of mutations. Thus, in order to formally study the sensitivity of our model to these factors we first introduce some new parameters. The parameters ν_E^ρ and ν_P^ρ, defined such that $\rho_E = \nu_E^\rho \rho_N$ and $\rho_P = \nu_P^\rho \rho_N$, are the proliferative advantages of the E and P populations over the N population, with minimum and maximum values of 1 and 1.5, i.e., no advantage and a 50% proliferation advantage. Similarly, we investigate the sensitivity of the model to invasive advantages afforded to the amplified sub-populations via the parameters ν_E^D and ν_P^D, defined analogously by $D_E = \nu_E^D D_N$ and $D_P = \nu_P^D D_N$. To investigate how the phylogenetic ordering of mutations influences the amplification pattern we see across our simulations in Section 3.4.2, we chose to introduce the EGFR and PDGFRA amplified sub-populations at various introduction times, t_E^* and t_P^* determined by the size, N_E or N_P, of the growing tumour, which we also include in the sensitivity analysis with minimum and maximum values of $3N_I$ and $9N_I$, where N_I is size of the initial population of non-amplified tumour cells.

In addition to these parameters, we include the introduction locations of the mutated populations, x_E^* and x_P^*, and the proliferation rate, ρ_N, and diffusion coefficient, D_N, of the non-amplified population of tumour cells (N) in our sensitivity analysis. As detailed in Section 3.3, we previously ran simulations for four (ρ_N, D_N) pairs to represent the heterogeneity of tumours present in our patient cohort and calculated the average proportions from these four simulations, while other parameters in the model were varied. Therefore, we assign each parameter uniform pdfs with minimum and maximum values of 3 and 30, with appropriate units, to represent tumours with varying degrees of proliferative and invasive capabilities.

The interactions, α_{PE} and α_{EP}, are the final parameters to be included in the

Table 3.1: Parameter definitions and the minimum and maximum values of their corresponding uniform distributions.

Parameter	Definition	(Min, Max)	Units
ν_E^ρ	proliferative advantage of E cells	(1, 1.5)	unitless
ν_P^ρ	proliferative advantage of P cells	(1, 1.5)	unitless
ν_E^D	invasive advantage of E cells	(1, 1.5)	unitless
ν_P^D	invasive advantage of P cells	(1, 1.5)	unitless
N_E	tumour size when E is introduced	$(3N_I, 9N_I)$	cells/mm^2
N_P	tumour size when P is introduced	$(3N_I, 9N_I)$	cells/mm^2
x_E^*	introduction location of E cells	$(x_c^* - 1.5, x_c^* + 1.5)$	mm
x_P^*	introduction location of P cells	$(x_c^* - 1.5, x_c^* + 1.5)$	mm
ρ_N	proliferation rate of N cells	(3, 30)	1/year
D_N	diffusion coefficient of N cells	(3, 30)	mm^2/year
α_{PE}	effect of P on E	(-5, 5)	unitless
α_{EP}	effect of E on P	(-5, 5)	unitless

sensitivity analysis. We note that we previously made the assumption that the interactions between the EGFR and PDGFRA amplified sub-populations were symmetric, that is $\alpha = \alpha_{PE} = \alpha_{EP}$. This assumption allowed us to reduce the number of parameters in our model and study the amplification patterns for the three key cases of competition ($\alpha = -5$), cooperation ($\alpha = 5$) and neutralism ($\alpha = 0$) throughout Section 3.4.2. It is, however, possible that these interactions are non-symmetric and one of the other six interaction types detailed in Table 2.1 could explain the amplification patterns observed in the biopsy data in Figure 3.1, or perhaps an interaction scenario where both populations are competing, but not to the same degree. Thus, in the following sensitivity analysis we allow α_{PE} and α_{EP} to take different values and sample them independently from uniform distributions with minima and maxima of -5 and 5.

Briefly, the first step of the sensitivity analysis is to conduct the LHS for which we choose a sample size of 2000; each of the 12 parameter distributions given in Table 3.1 are divided into 2000 intervals of equal probability and a sample is drawn from each. These samples are then randomly grouped and 2000 simulations are run, from which we record the four outputs of interest: the proportions of neither, only EGFR, only PDGFRA and both genes amplified. The inputs and outputs are then rank transformed and PRCC values are calculated between each parameter and each output of interest, with values ranging from -1 to $+1$. The sign and magnitude of the PRCC values indicate the qualitative relationship between the input and output variables and the importance of parameter uncertainties in accurate prediction of the model outputs [10]. A significance test is

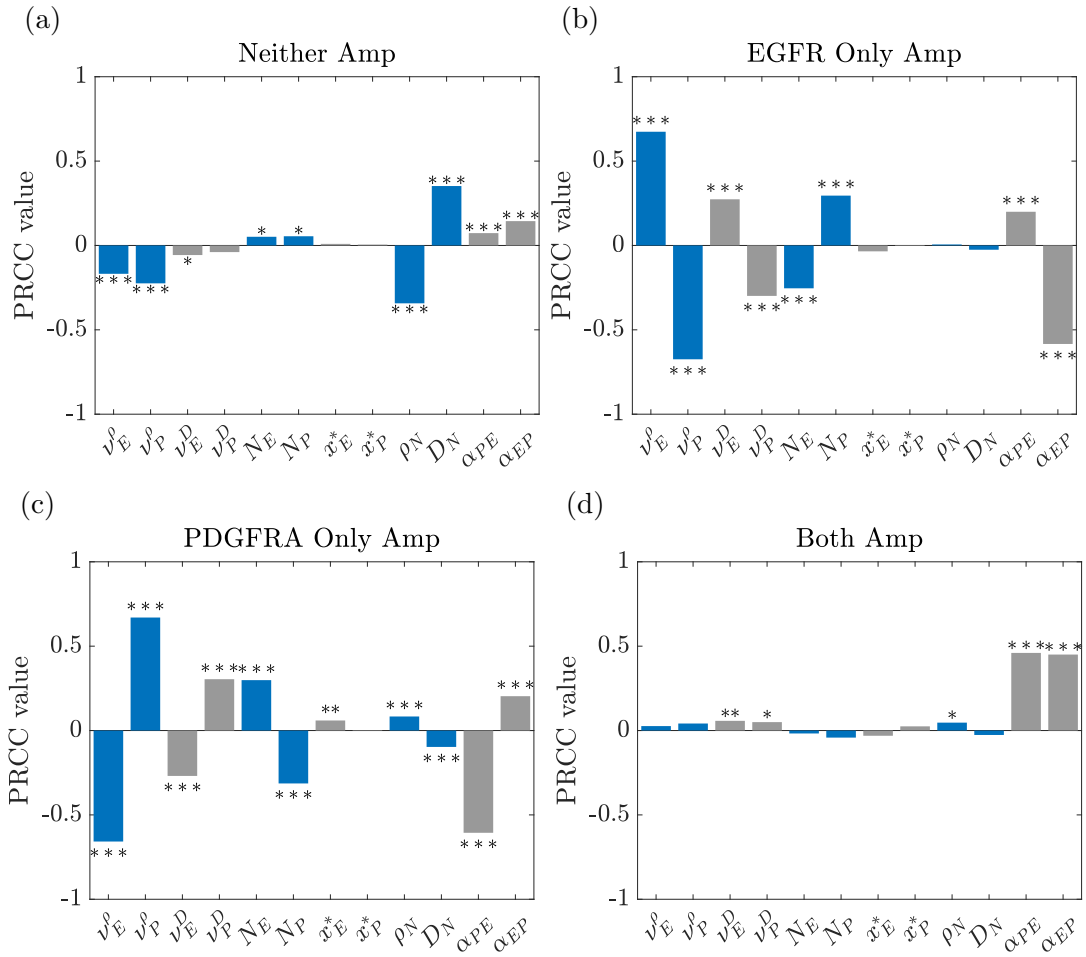


Figure 3.7: Bar plots showing PRCC values between each model parameter detailed in Table 3.1 and the four outputs of interest: the proportion of simulations with (a) neither, (b) only EGFR, (c) only PDGFRA and (d) both genes amplified. PRCC values significantly different from zero at the 0.05 (*), the 0.01 (**) and the 0.001 (***) levels are highlighted.

conducted to test whether each PRCC value is different from zero, thus indicating any significant correlations between outputs and parameters. A more in depth description of the LHS-PRCC protocol is provided in [10] and [60]. We implemented the sensitivity analysis utilising code by Massey et al. [62], which can be found at `https://github.com/scmassey/model-sensitivity-analysis` along with further details of the LHS-PRCC method.

The results from the LHS-PRCC sensitivity analysis for the four outputs of interest from our model—given by Eq.s (2.1)–(2.3), (2.7) and (3.1)–(3.6) and initial conditions (3.7)—and the 12 parameters detailed in Table 3.1 are shown in Figure 3.7. As expected from our analysis in Section 3.4.2, the proportions of simulations with only EGFR and only PDGFRA amplified are both strongly correlated to selection advantages and the timings of mutations, with the effects being reflected for each output. For example, there is a strong positive correlation between the parameter ν_E^ρ, which affords the EGFR amplified population a proliferative advantage, and the proportion of simulations with only EGFR amplified, whereas there is a strong negative correlation between this parameter and the proportion with only PDGFRA amplified. Affording these amplified populations invasive advantages, through ν_E^D and ν_P^D, and introducing them at earlier times, $\tilde{t}_E$ and $\tilde{t}_P$, in the growing tumour also has a similar, albeit slightly weaker, effect on the proportions of simulations with only one of the genes amplified observed.

In Figure 3.7a, we observe that the proportion of simulations with neither gene amplified is most sensitive to the parameters ρ_N and D_N. This may be due to the effects at the edges of the simulated tumours, where the total tumour cell population, $T = N + E + P$, is above the threshold of detection and the amplified populations, E and P, remain undetected below the threshold (as illustrated in Figure 3.3a), so these points in the simulations contribute to the proportion of the simulation with neither gene amplified. As the parameter ρ_N decreases and D_N increases, the profile of the tumour edges becomes flatter and this area at the edge of the tumour becomes wider, thus contributing more to the proportion of the simulated tumour with neither gene amplified. This suggests that to represent amplification patterns for a population of heterogeneous GBMs it is important to incorporate this heterogeneity into our mathematical modelling by considering a range of ρ_N and D_N values, as we did throughout Section 3.4.2.

In Figure 3.7d, we notice that only the interaction parameters, α_{PE} and α_{EP}, have a strong impact on the proportion of simulations with both genes amplified; this is, again, consistent with our previous observations in Section 3.4.2, where we saw that the proportion of simulations with both genes amplified increased as we

moved from the competitive case through to the neutral and cooperative cases. The other proportions in the simulations are also strongly correlated to these parameters. We observe that if α_{PE} and α_{EP} have opposite signs, say $\alpha_{PE} > 0$ and $\alpha_{EP} < 0$, then the EGFR amplified population will benefit, while the proportion with only PDGFRA amplified will decrease. The case where both parameters have the same sign is more difficult to understand due to the competing effects; a positive α_{PE} will increase the only EGFR Amp proportion and have a stronger negative effect on the only PDGFRA Amp proportion, whereas a positive α_{EP}, will have the opposite effect. Since the negative effects are stronger in this case, it is likely that the proportions with only one gene amplified will decrease, as the proportion with both amplified will increase when both interaction parameters are positive. Meanwhile, if α_{PE} and α_{EP} are both negative, the proportion with both genes amplified will decrease, as the EGFR and PDGFRA amplified sub-populations compete with one another; this is likely to result in each amplified population occupying distinct regions of the tumour or one population dominating, depending on the strength of the competitive interactions and various other factors, such as one population being introduced and becoming established before the other.

Finally, we note that the PRCC values between the introduction locations and each of the four outputs of interest are close to zero, indicating that they are not strongly correlated. While this may be expected, since our results in Figure 3.6 in Section 3.4.2 and Figures C.4 and C.5 in the Appendix C do not show that changing the introduction locations has a big effect on the amplification patterns observed, it is also possible that this is due to a non-monotonic relationship between the introduction locations and the outputs. In this case, it is possible to remove the non-monotonicity from the problem by dividing the domains of x_E^* and x_P^* into two; this process is discussed in further detail in Appendix D, where the results of a second LHS-PRCC analysis with the non-monotonicities accounted for can be found in Figures D.1 and D.2. Upon removing the non-monotonicity, we find, however, that there are still no very strong correlations between the introduction locations of the amplified sub-populations and the outputs of interest, while the results for the other 10 model parameters remain consistent with the results presented here.

3.5 Discussion

In this chapter, we have used a novel mathematical model describing the co-evolution of three distinct tumour cell sub-populations to investigate the nature of interactions between cells with two common mutations occurring in GBMs, namely amplification of the genes encoding the EGFR and PDGFRA proteins. We conducted an *in silico* investigation into the levels of EGFR and PDGFRA amplified sub-populations observed under competitive, cooperative and neutral interactions between these cell types and compared our results to population levels of amplification observed in image-localized biopsy data from an initial cohort of GBM patients, where a high proportion of biopsies had only the EGFR gene amplified, a smaller proportion had both or neither gene amplified and very few showed amplification of only the PDGFRA gene (see Figure 3.1b).

In carrying out computational simulations, we found that the amplification patterns observed in simulated tumours under each of the different interaction assumptions did not match those observed across the patient biopsy data when we assumed that cells with the EGFR and PDGFRA genes amplified did not differ in their dynamics, shown in Figure 3.3b. We note that this was to be expected and is consistent with research suggesting dynamical differences between these sub-populations. For example, a study reconstructing the phylogeny of GBMs suggests that EGFR amplification is a mutation that arises earlier than amplification of the PDGFRA gene [108], while other studies have found that tumours with EGFR amplified are more invasive, which could indicate that EGFR amplified cells themselves are more invasive [81, 124]. Following this, we explored the effects of introducing differences between the sub-populations in our simulations through changing various model parameters. Since a high proportion of biopsies across the patient data have the EGFR gene amplified but no amplification of the PDGFRA gene, we investigated the effects of giving EGFR amplified cells various selection advantages over PDGFRA and non-amplified cell types, shown in Figure 3.4 and Figure C.1 in Appendix C. While these simulations do not achieve the same amplification patterns observed in the biopsy data, affording EGFR amplified cells a selection advantage does help to produce the high levels of EGFR amplification required, particularly in the competition and neutral interaction cases. We note that further investigation of other degrees of invasive and proliferative advantages may improve the results we see; throughout Section 3.4.2 we only looked at quite large advantages of 50% of the respective parameters, allowing different degrees of advantages in the model is something that will be addressed in Chapter 4. We then chose to investigate the effects of changing

the phylogenetic ordering of mutations and found that introducing the EGFR amplified sub-population earlier in the evolution of our simulated tumours also helped to produce the desired higher level of EGFR amplification under competitive and neutral interaction assumptions, whereas the proportion of points with both genes amplified was much higher than observed in the biopsy data in the cooperative case. Finally, we looked at introducing the EGFR and PDGFRA amplified sub-populations in different locations in the growing tumour simulations. We found that changing these locations did not have such a large effect on the amplification patterns observed as the selection advantages and phylogeny did, however it could be an important factor in replicating amplification patterns observed in individual GBMs with distinct regions of EGFR and PDGFRA amplification.

We also conducted a LHS-PRCC sensitivity analysis, which highlighted the sensitivity of the amplification patterns observed across simulations to the selective advantages and introduction times of the amplified sub-populations and the type of interactions occurring between them. Consistent with our finding in Section 3.4.2, this analysis also highlighted that the amplification patterns were not strongly correlated to the introduction locations; while this may be the case, the weak correlation observed could be a result of only considering a small range of tumour locations as a result of the small size of the tumour at early introduction times providing little scope for spatial variation. This is a factor that could be investigated in future work.

Although the simulation results presented in this chapter do not perfectly match the patterns of EGFR and PDGFRA amplification observed across the patient biopsy data in Fig. 3.1b, our *in silico* modelling approach has allowed us to investigate how different interaction assumptions influence the amplification patterns in simulated tumours and explore the effects of changing the parameters and the timing and position of sub-population introductions in our model. We found that some of these changes improved our simulation results with respect to the biopsy data and were also consistent with suggestions about EGFR amplified sub-populations found in the literature.

In Section 3.4.2, we only explored each factor individually, whereas a combination of factors relating to the selection advantages of EGFR and PDGFRA amplified cells as well as the timing and location of mutations is likely to be influencing the amplification patterns across the biopsy data. This combination of factors makes it difficult, at this stage, to deduce the nature of interactions between cell types that is driving these patterns, however, this will be addressed

in Chapter 4, where we build on the work presented in this chapter. Across our simulation results, nonetheless, we did find that the competitive and neutral cases approximated the patterns observed in the biopsy data better than the cooperative case, suggesting that EGFR and PDGFRA amplified sub-populations are not strongly cooperating with one another.

Throughout our investigation in Section 3.4.2, we only studied one strength of competitive and cooperative interactions between EGFR and PDGFRA amplified sub-populations and also assumed that these interactions were symmetric, whereas these cells may be interacting differently and to stronger or weaker degrees. To shed light on this, in Section 3.4.3, we allowed α_{PE} and α_{EP} to take different values. We found that, if they have opposite signs, one amplified population will benefit while the other decreases and that if they have the same sign but are not equal the results are more difficult to interpret due to competing effects. The effect of this on the amplification patterns observed across a cohort of tumours requires further investigation, which is something that may be studied in future work. In Chapter 4, however, we remove the assumption that the interactions between EGFR and PDGFRA amplified sub-populations are symmetric and explore different strengths of cooperation and competition in our work.

Determining the nature of interactions between EGFR and PDGFRA amplified sub-populations in GBMs is a complex biological problem, with factors relating to selection advantages and the phylogeny of these tumours influencing the balance of populations we see in a tumour, as we have demonstrated with our *in silico* investigation in this chapter. To be able to untangle the influence from and the nature of interactions from the effects of these other factors, more data may be required. In this study, we had biopsy data for a cohort of 25 patients and a larger number may enable us to gain more insight into the pattern of EGFR and PDGFRA amplification. The patient cohort data is expanded in Chapter 4, where we use an inference algorithm to infer the model parameters to see if any further insight into the interactions between these amplified sub-populations in GBMs.

Chapter 4

Parameter Inference to characterise EGFR and PDGFRA amplified glioblastoma sub-populations

4.1 Introduction

The work presented in this chapter follows on from that detailed in Chapter 3. Here, we build on the knowledge gained from the *in silico* investigations carried out to infer estimates of the dynamics and interactions of EGFR and PDGFRA amplified sub-populations in a final set of patient biopsy data.

This chapter begins by presenting tissue analysis data in Section 4.2, building on the preliminary dataset presented in Chapter 3. We briefly discuss the need to modify the copy number aberration (CNA) threshold previously used to determine the amplification status of the biopsies, before illustrating a patient example and providing the details of the tissue samples from the final cohort of patients included in the study.

Next, we discuss the need to introduce stochasticity into the model, giving details of the approach taken in Section 4.3, before then moving on to parameter inference work in Section 4.4. We detail the inference algorithm used and test whether the parameters in our model can be inferred for example synthetic datasets. Finally, in Section 4.5, the model parameters are inferred for the patient biopsy data and the results are discussed.

4.2 Image-localised biopsies and tissue analysis from a cohort of GBM patients

Following on from the work in Chapter 3, an additional 30 patients were recruited to the study by researchers at the Mayo Clinic, taking the total patient cohort to 55 patients. As previously described in Section 3.2, patients with clinically suspected GBM undergoing preoperative MRI for surgical resection were recruited to the study and the absence of previous treatment was confirmed. Institutional review board approval was obtained, along with written and informed consent from each participant prior to being enrolled. Multiple image-localised biopsies were collected during surgery from each patient; more detail of the biopsy collection protocol can be found in Section 3.2 and [41].

Copy number aberration (CNA) values associated with the EGFR and PDGFRA genes are determined for the biopsy samples using array comparative genomic hybridization (aCGH) or whole exome sequencing, more details of these processes can be found in [40] and [131]. Each biopsy is then classified as being amplified in each gene based on these CNA values, with a value above a certain threshold corresponding to amplification and vice versa. Previously, we chose to use a CNA value of 2.2 for this threshold. This threshold was chosen so that it would be sensitive to densities of EGFR and PDGFRA amplified cell populations as low as 10% in the tissue samples; this was based on the knowledge that each biopsy is likely to contain a mixture of amplified and non-amplified cells in each gene and that all cells not amplified in each gene will have a CNA equal to 2 and the assumption that all cells amplified in each gene would have a CNA value of 4.

In the updated dataset, however, only rounded, whole number CNA values were available for a subset of the patient biopsies. Thus, we choose to use a CNA threshold of 2.5 to determine the amplification status of the biopsies as this threshold can be used consistently across the whole dataset. In this way, using the same knowledge and assumptions about the CNA values of amplified and non-amplified cell populations, the amplification threshold will now be sensitive to densities of amplified cell populations as low as 25% of the tissue sample.

Analogously to Chapter 3, the CNA threshold of 2.5 is calculated as follows:

$$\begin{aligned}\text{CNA value of tissue sample} &= \begin{pmatrix}\text{Non-amp fraction of sample} \times \text{CNA value of non-amp cells}\end{pmatrix} \\ &\quad + \begin{pmatrix}\text{Amp fraction of sample} \times \text{CNA value of amp cells}\end{pmatrix} \\ &= (0.75 \times 2) + (0.25 \times 4) \\ &= 2.5.\end{aligned}$$

Intratumoural heterogeneity: a patient example

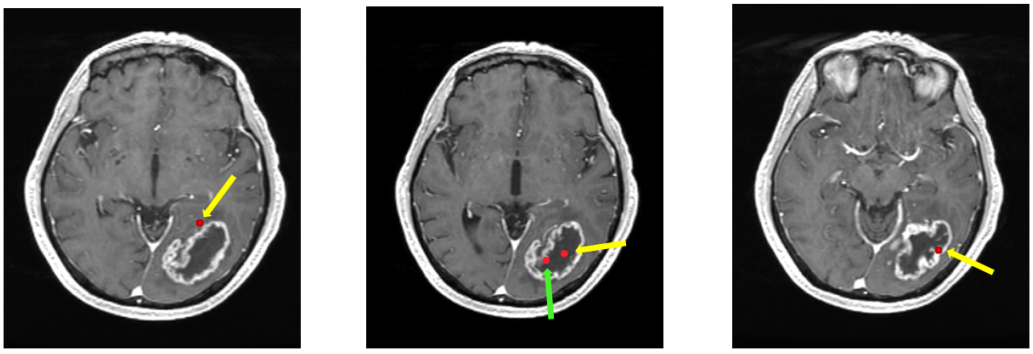


Figure 4.1: Three axial T1Gd MRI slices of a patient with a clinically diagnosed GBM to the rear of the brain in the left occipital lobe. The location of a biopsy sample as recorded during surgery is shown on each slice by the red dot and highlighted by an arrow.

A patient with clinically suspected GBM who was included in this study underwent surgical resection of their GBM, during which 8 biopsy samples were taken. The surgical team recorded the locations from which each of these biopsies were sampled, which were then co-registered with the individual's preoperative MRI scans. In this way, each biopsy has a set of (x, y, z) coordinates matching the location on the MRI scan from where it was sampled, with the z coordinate corresponding to the slice number.

Figure 4.1 shows 3 axial slices of the patient's preoperative T1Gd MRIs. On each of these slices, the (x, y) coordinates of any biopsy samples with the matching z coordinate are marked by a red dot and highlighted by an arrow. Of the 8 sampled biopsies, 4 are shown on these 3 slices, while the remaining 4

biopsies have coordinates on 4 different MRI slices that are not shown here for brevity.

It can be observed from the images shown in Fig. 4.1 that the neurosurgeon has sampled biopsies from a variety of regions within the tumour: (a) the biopsy location is recorded just outside the enhancing tumour region, where the invasive edge of the tumour is likely to be; (b) two biopsy locations are recorded on this slice, with one being close to the centre of the necrotic tumour core (yellow arrow), while the other is on the inside edge of the enhancing lesion (green arrow); (c) the biopsy location is within the enhancing region of the tumour. In this way, an insight into the genomic profile of these different regions of the tumour is gained when the samples are sent for analysis.

CNA values are determined for each of the biopsies, where it is found that this patient's tumour exhibits heterogeneity in EGFR and PDGFRA amplification. For example, of the biopsies shown in Fig. 4.1, two are amplified in only EGFR; the biopsies sampled from the locations shown in (a) and (c). Meanwhile, the biopsy from the location highlighted by the green arrow in (b) is amplified in neither gene, while the other biopsy shown on the same slice is amplified in both genes. The location and amplification status of each of the eight biopsies sampled from this individual's tumour are summarised in Figures 4.2(a)-(d). These figures show the locations of the eight biopsies and their distribution throughout the various tumour regions. Here it can be seen that the neurosurgeon has sampled at least one biopsy from the invasive margin, the T1Gd enhancing tumour and the necrotic tumour core as segmented from the preoperative MRI scans. Of these eight biopsies, two are amplified in neither of the genes, five are amplified in only the EGFR gene, while one is amplified in both the EGFR and PDGFRA genes.

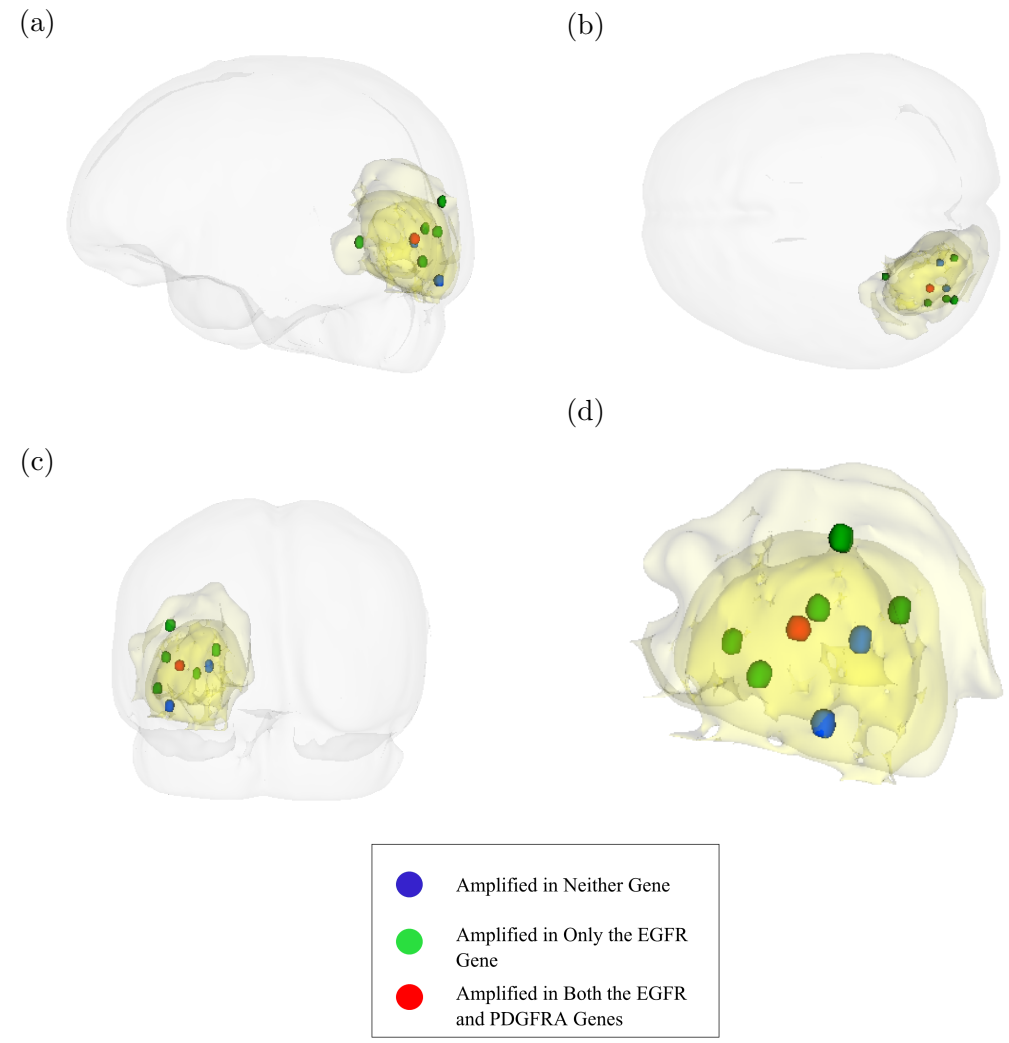


Figure 4.2: Images showing biopsy sampling locations and their corresponding amplification status for an individual with a clinically diagnosed glioblastoma. The brain domain segmented from the patient's MRI scans is shown pale grey in (a) sagittal, (b) axial and (c) coronal planes. A lesion is shown in pale yellow to the rear of the brain in the left occipital lobe. Three different layers to this lesion are shown: the outermost layer is the segmented volume from T2 MRI scans representing tumour edema; the middle layer is the segmented contrast enhancing volume from T1Gd MRIs, corresponding to the densest part of the tumour; and the innermost layer is the non-enhancing tumour core segmented from T1Gd MRIs, showing the necrotic core of the tumour. A closer image of this tumour lesion is shown in (d) without the surrounding brain domain. Within this tumour lesion small volumes are shown at the locations where biopsies were sampled as recorded by the surgical team during surgery. The colour of these volumes represent the amplification status of each biopsy as determined through tissue analysis. The brain domain and tumour volumes were segmented by a member of the Mathematical NeuroOncology Lab at Mayo Clinic, Phoenix, Arizona. These images were created using the open source software 3D Slicer (version 4.10.2).

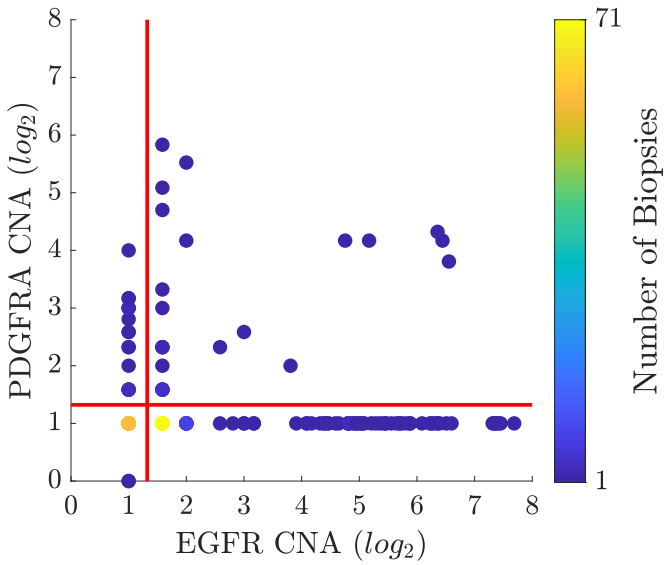


Figure 4.3: Scatter plot showing the EGFR and PDGFRA CNA whole number values for each of the 240 patient biopsies. The values are plotted on a log_2 scale and the colour of each point represents the number of biopsies for each CNA value pair. The red lines show the thresholds used to determine amplification status for each of the EGFR and PDGFRA genes, that is a CNA value of 2.5 in each case.

Sampling from a cohort of patients with GBM

A total of 240 biopsies that contained adequate tumour and/or DNA content to undergo analysis were collected from 55 patients with clinically suspected primary GBM; we note that this data is also published in [131]. EGFR and PDGFRA CNA values were successfully determined for these biopsies through aCGH analysis or whole exome sequencing. The number of biopsies sampled from each patient ranged from 1-10, with a median of 4 samples coming from an individual. Only rounded whole number CNA values were available for the samples from 24 patients (116 biopsies), as opposed to values with 6 decimal points for other samples. Therefore, we chose to use rounded whole number CNA values for all samples in order to be consistent across our entire dataset, which are shown in the scatter plot in Fig. 4.3 along with the thresholds used to determine amplification status in each of the EGFR and PDGFRA genes.

EGFR amplification was the more commonly observed genetic alteration, with 162/240 samples having a CNA value associated with EGFR amplification, whereas 40/240 were determined to be PDGFRA amplified. Of these amplified samples, 24 were found to have amplification of both the EGFR and PDGFRA genes.

The data shown in Fig. 4.3 highlight the heterogeneity in CNA values ob-

served across samples from this cohort of patients, where biopsies with up to 206 copies of the EGFR gene and 57 copies of the PDGFRA gene were found. Even within individual tumours, large variations in CNA values are observed; for example, of the four biopsies taken from one patient's tumour, two did not exhibit amplification in either gene and two were highly amplified in the PDGFRA gene, one sample having 46 copies and the other 34. This highlights the importance of multi-region sampling from GBM tumours, as it demonstrates that the genetic profile of one region of a tumour can be very different to another, which may have implications for therapeutic decisions and treatment outcomes.

For each patient, we then determined the proportion of their biopsies that were found to be amplified in neither gene, only the EGFR gene, only the PDGFRA gene and, finally, both of the EGFR and PDGFRA genes simultaneously. The proportions calculated for each of the 55 patients are summarised as a box plot in Fig. 4.4a and the means of these proportions are shown as a spider plot in Fig. 4.4b. Figure 4.4a highlights the heterogeneity of amplification patterns observed across the patient cohort, with some patients having all their biopsies amplified in neither, only one or both of the genes. Meanwhile, other patients have differing amplification among their biopsies, as illustrated by the patient example discussed earlier in this section. The spider plot shown in Fig. 4.4b demonstrates that the highest mean proportion of biopsies are those amplified in EGFR, followed by amplification in neither gene; the mean proportions of biopsies that are amplified in only PDGFR or both genes are both at low levels. We note that the mean proportion of biopsies amplified in neither gene is slightly higher in Fig. 4.4b than in Fig. 3.1b, while the mean amplified proportions are slightly lower; this is likely to be a result of the slightly different CNA thresholds used to determine amplification in this work and in the preliminary dataset in Chapter 3.

Of the 55 tumours for which copy number data for at least one biopsy were determined in this dataset, 12 had at least one biopsy sampled where both the EGFR and PDGFRA genes were amplified. Meanwhile, 45/55 (81%) and 16/55 (29%) tumours were found to have at least one biopsy amplified in the EGFR and PDGFRA genes, respectively. We note that these numbers are much higher than those observed in data from the Cancer Genome Atlas (TCGA) [125], where 41% and 10% of the 206 tumours analysed were found to be amplified in EGFR and PDGFRA, respectively, and 5 cases were observed to have co-amplification of these genes [107]. The numbers of tumours exhibiting amplification of these genes is likely to be much higher in our dataset than in TCGA data due to the fact that

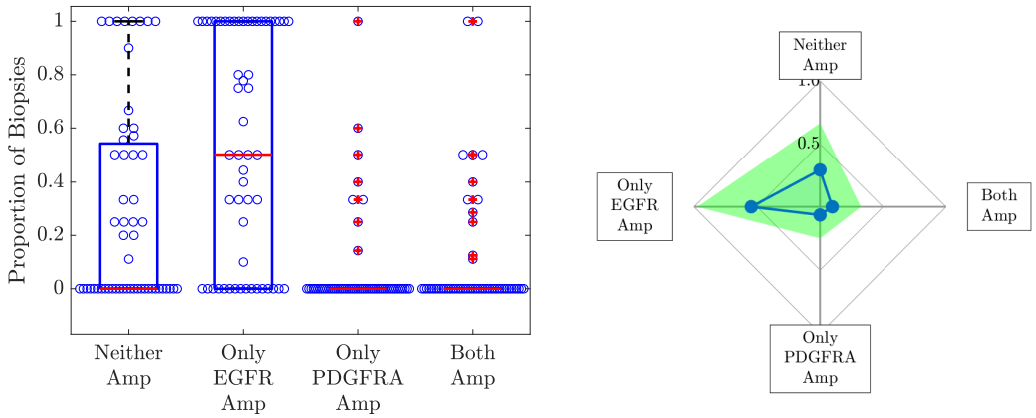


Figure 4.4: (a) A box plot summarising the proportion of each individual's biopsies that were determined to be amplified in neither gene (Neither), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both of the EGFR and PDGFRA genes (Both Amp) for the 55 patients. Each of the blue circles overlaid on the box plot represents the relevant proportion of an individual's biopsies for each category. The means of these proportions across the 55 patients are shown in (b).

multiple biopsies are sampled from various regions of each tumour in our dataset, whereas single samples are analysed for each case in the TCGA dataset. Thus, where only one biopsy is sampled, amplification may not be identified in cases where the tumour is not amplified everywhere, whereas multi-region sampling is more likely to identify these cases.

4.3 Introducing stochasticity in a simulated population of GBMs

The data from our patient cohort, illustrated in Fig. 4.4a, shows that the proportions of biopsies amplified in the EGFR and PDGFRA genes vary from patient to patient, with some patients having all their biopsies without any amplification in these genes, others having all biopsies with one or both genes amplified and others somewhere in between with a mixture of amplified and non-amplified biopsies. Natural variation and the fact that mutations occur randomly are likely to play a role in contributing to this observed variation and so, to replicate this heterogeneity arising in our patient cohort, we introduce stochasticity to some of the model parameters and run a number of simulations, M, to create a simulated population of GBMs.

As the biopsy data are calculated as the mean from a cohort of 55 individual

tumours, where each tumour varies in their proliferative and invasive ability, we use a variety of ρ_N and D_N pairs to reflect this. (Recalling that the proliferation and invasion rates of the EGFR amplified sub-population are defined as $\rho_E = \nu_E^\rho \rho_N$ and $D_E = \nu_E^D D_N$, with the parameters for the PDGFRA amplified cell population defined analogously, the proliferative and invasive abilities of these amplified sub-populations are also varied implicitly.) In Chapter 3 and [68], this variability was accounted for by producing four simulations using two values for each of ρ_N and D_N to mirror the range of parameters observed in unpublished patient databases; to represent high parameter values we used $\rho_N = 30$/year and $D_N = 30\text{mm}^2$/year and for low values we used $\rho_N = 3$/year and $D_N = 3\text{mm}^2$/year as in [36]. Instead, we now make the assumption that the proliferation and invasion parameters are uniformly distributed over these ranges and, thus, in each of our M simulations, ρ_N and D_N are selected from the distributions

$$\rho_N \sim U(3, 30)/\text{year} \quad \text{and} \quad D_N \sim U(3, 30)\text{mm}^2/\text{year}.$$

In the previous chapter, the effect of changing the location of mutations, x_E^* and x_P^*, on the amplification patterns observed in our simulated tumours was explored. While the model was not found to be particularly sensitive to these parameters, they may still play a role in contributing to the observed heterogeneity of amplification levels in Fig. 4.4a. Since the mutations leading to the introduction of amplified sub-populations occur randomly during cell proliferation, we now assume that the locations at which these populations are introduced in the simulations are also random. Further, we make the assumption that these mutations are more likely to occur where more proliferation is taking place and less likely to occur where fewer cells are replicating. Since we previously assumed that E and P cells mutate only from N cells, we determine this likelihood based on the proliferation of the N population of cells. Therefore, at the time point, t_E^*, at which the EGFR amplified population is introduced in each simulation, the location x_E^* is randomly sampled from the set of simulation spatial mesh points x_i with weights $w_{x,i}$ defined by

$$w_{x,i} = \frac{f_N(x_i, t_E^*)}{\Sigma_j f_N(x_j, t_E^*)}, \tag{4.1}$$

where f_N is given by Eq. (2.7). Thus, the spatial mesh points where the net proliferation of the N population is highest at this time point are more likely to be selected as the introduction location of population E. Similarly, x_P^* is randomly sampled at time t_P^* with analogously defined weights.

In Chapter 3, we explored the effect of changing the introduction times of the two amplified sub-populations on the mean proportions of amplified simulations observed. These introduction times, t_E^* and t_P^*, are defined as the first time point in a simulation after the growing population of N tumour cells had reached a particular size, N_E or N_P. By choosing a range of values for N_E and N_P, we explored the effect of changing the phylogenetic ordering of mutations and found that introducing the EGFR amplified sub-population earlier in the evolution of our simulated tumours also helped to produce the desired higher level of EGFR amplification under competitive and neutral interaction assumptions.

Since the mutations leading to the introduction of these amplified sub-populations are random events, we now assume that the parameters N_E and N_P are random variables with an associated distribution and mean that may differ between the EGFR and PDGFRA amplified sub-populations. In this work, we choose to model them with Gamma distributions,

$$N_E \sim N_I + \Gamma(k_E, \theta_E) \quad \text{and} \quad N_P \sim N_I + \Gamma(k_P, \theta_P),$$

where $k_E, k_P > 0$ are the shape and $\theta_E, \theta_P > 0$ are the scale parameters. N_E and N_P are modelled in this way so that the possible range of values are $[N_I, \infty)$, with means $N_I + k_E\theta_E$ and $N_I + k_P\theta_P$, respectively. We recall that N_I is the size of the initial population of N cells, as defined in Chapter 3.

As mentioned previously, a mutation event occurring in a cell leads to the creation of a single EGFR or PDGFRA amplified cell and, while there are likely to be many such events occurring during the growth of a GBM, we assume that each of the EGFR and PDGFRA amplified sub-populations only become established within the tumour at most once, as in Chapter 3 and [68]. Thus, the scale parameters, θ_E and θ_P, of the gamma distribution are the change in size of the tumour between mutation events leading to the creation of an EGFR or PDGFRA amplified cell; and the shape parameters, k_E and k_P, are the average number of such events occurring before the amplified sub-population is introduced into the growing tumour and has the opportunity to become established. Therefore, a small value of k_i means that few mutations occur before population i is introduced into the growing tumour and a small value of θ_i means these mutations events are occurring regularly as the tumour grows.

Modelling N_E and N_P in this way introduces an additional four unknown parameters into our model and, thus, the final list of unknown model parameters to be inferred are given in Table 4.1.

Table 4.1: Definitions of parameters to be inferred.

Parameter	Definition	Units
ν_E^ρ	proliferative advantage of E cells	unitless
ν_P^ρ	proliferative advantage of P cells	unitless
ν_E^D	invasive advantage of E cells	unitless
ν_P^D	invasive advantage of P cells	unitless
k_E	shape parameter of gamma distribution for N_E	unitless
θ_E	scale parameter of gamma distribution for N_E	cells/mm^2
k_P	shape parameter of gamma distribution for N_P	unitless
θ_P	scale parameter of gamma distribution for N_P	cells/mm^2
α_{PE}	effect of P on E	unitless
α_{EP}	effect of E on P	unitless

4.3.1 Defining a simulated cohort of GBMs

Throughout the following parameter inference work, we produce a simulated cohort of GBMs for a given candidate parameter vector, $\eta^* = (\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP})$. Given η^*, we define a *simulated cohort of M GBMs* as $M \in \mathbb{N}$ simulations of our model, given by Eq.s (2.1)–(2.3), (2.7) and (3.1)–(3.6) and initial conditions (3.7), where each of the M model simulations are produced with the fixed parameters $\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, \alpha_{PE}, \alpha_{EP}$, while the stochastic parameters ρ_N, D_N, x_E^* and x_P^* are sampled as described in the previous section and the parameters N_E and N_P are sampled from the Gamma distributions,

$$N_E \sim N_I + \Gamma(k_E, \theta_E) \text{ cells/mm}^2, \tag{4.2}$$

$$N_P \sim N_I + \Gamma(k_P, \theta_P) \text{ cells/mm}^2, \tag{4.3}$$

for each of the M simulation runs.

Each simulation of the model is run to a biologically relevant size, which is chosen to represent a typical size of a GBM tumour at the time of diagnosis. Thus, as described in Chapter 3, we run each simulation until the width of the total tumour cell population above the threshold of $0.8K$ in each simulation is 36.2mm, to reflect that the average tumour diameter at diagnosis is 36.2 mm (unpublished patient data); more detail of this process can be found in the previous chapter of this thesis. In this way, a sample of simulated GBMs is generated for the given parameter vector, η^*, that reflects the heterogeneity observed in the GBMs sampled from the patient cohort detailed in Fig. 4.4a in Section 4.2.

For each of the individual simulated GBMs, the proportion of the tumour

with both of the EGFR and PDGFRA genes amplified, only one gene amplified and neither of the genes amplified is then calculated. This process in analogous to that outlined in Chapter 3, but taking into account the higher CNA threshold used to define amplified biopsies in the updated set of patient data. As the new patient data are for the proportions of biopsies containing EGFR and PDGFRA amplified cells above a given density threshold, now assumed to be 25% of the tissue sample instead of 10% as previously, we modify the equivalent measure for each simulated tumour defined in Chapter 3 accordingly. For example, the proportion of the simulated tumour with only EGFR amplified at time t, $A_E(t)$, is now calculated as

$$A_E(t) = \frac{\int_0^L H(E(x,t) - 0.25K)H(0.25K - P(x,t))\,\mathrm{d}x}{\int_0^L H(E(x,t) + P(x,t) + N(x,t) - 0.25K)\,\mathrm{d}x},$$
$$\approx \frac{\text{Number of mesh points with } E > 0.25K \text{ and } P < 0.25K \text{ at time } t}{\text{Number of mesh points with } T > 0.25K \text{ at time } t}, \tag{4.4}$$

where $T = E+P+N$ is the total tumour cell population and $H(\cdot)$ is the Heaviside step function. The proportions with neither gene, only the PDGFRA gene and both genes amplified, $A_N(t)$, $A_P(t)$ and $A_B(t)$, are defined and calculated in a similar way. Each of these measures is then calculated at the time when each simulation reaches the biologically relevant size, defined as having $T > 0.8K$ at a width of 36.2mm. This process is analogous to the approach taken in Chapter 3, where it is illustrated schematically in Fig. 3.3a.

Throughout the following parameter inference work, these proportions calculated from a simulated cohort of GBMs for each candidate parameter vector are then compared to the proportions of amplified and non-amplified biopsies in the patient cohort data. To do this, we make the assumption that the proportion of biopsies amplified in each of the EGFR and PDGFRA genes for each patient is representative of the amplified proportions of their entire tumour. Further details of the inference methods we use and the metric used to compare the simulated and patient cohorts are discussed later in this chapter.

For the remainder of this thesis, each of the simulation runs of the model is produced using MatLab R2017a to implement a finite difference scheme in space with a uniform mesh size of 0.5 mm and a forward Euler time step of 0.001 years. As in Chapter 3, we consider a domain length of L = 200 mm. Since we are only interested in running simulations to a biologically relevant size and all

populations are introduced close to its centre, this domain is sufficiently large that tumour growth remains far from the boundaries, hence avoiding boundary condition artefacts.

4.4 Parameter Inference

In this work, we now employ an approximate Bayesian computation method based on sequential Monte Carlo, termed an ABC-SMC method [127], to estimate the unknown parameters in our model. First, we estimate the parameters for synthetic datasets, before then applying the inference scheme to the patient dataset detailed in Section 4.2.

4.4.1 ABC-SMC Algorithm

The ABC-SMC parameter inference algorithm requires a prior distribution, $\pi(\eta)$, to be proposed for the unknown parameters. From this prior distribution, a number of particles, $\{\eta^{(1)}, ..., \eta^{(R)}\}$, is then sampled and propagated through a sequence of intermediate populations, gradually evolving until they represent a sample population from the target posterior. In this work we employ the ABC-SMC algorithm as developed by Toni [127].

The ABC-SMC algorithm requires the user to predefine a number of inputs, namely: the number of populations, $Q+1$; the vector of tolerances, $\epsilon = (\epsilon_1, ..., \epsilon_Q)$; the perturbation kernels, K_q; the prior distribution, $\pi(\eta)$; the number of particles to sample R; and the function, $d(y^*, y_0)$, to determine the distance of the simulated data, y^*, from the data points, y_0. The ABC-SMC algorithm then proceeds as follows [127, 128]:

1: Initialise the tolerance vector, $(\epsilon_1, ..., \epsilon_Q)$; these are chosen such that $\epsilon_1 > ... > \epsilon_Q \geq 0$.
2: Set the population indicator $q = 0$.
3: Set the particle indicator $i = 1$.
4: **while** $q \leq Q$ **do**
5: **if** $q = 0$ **then**
6: **while** $i \leq R$ **do**
7: Sample η^* independently from $\pi(\eta)$.
8: Set $\eta_q^{(i)} = \eta^*$.
9: Set the weight for the particle $\eta_q^{(i)}$, $w_q^{(i)} = 1$.
10: Update the particle indicator, $i \leftarrow i + 1$.

end while
else
while $i \leq R$ **do**
Sample η^* from the previous population $\{\eta_{q-1}^{(1)}, ..., \eta_{q-1}^{(R)}\}$ with weights w_{q-1}.
Perturb the particle to obtain $\eta^{**} \sim K_q(\eta|\eta^*)$.
if $\pi(\eta^{**}) > 0$ **then**
Simulate a candidate dataset $y^* \sim f(y|\eta^{**})$.
Calculate the distance, $d(y^*, y_0)$.
if $d(y^*, y_0) < \epsilon_q$ **then**
Set $\eta_q^{(i)} = \eta^{**}$.
Calculate the weight for the particle $\eta_q^{(i)}$,
$$w_q^{(i)} = \frac{\pi(\eta_q^{(i)})}{\sum_{j=1}^{R} w_{q-1}^{(j)} K_q(\eta_q^{(i)}|\eta_{q-1}^{(j)})}$$
Update the particle indicator, $i \leftarrow i + 1$.
end if
end if
end while
end if
Normalise the weights.
Update the population indicator, $q \leftarrow q + 1$.
Reset the particle indicator, $i = 1$.
end while

In this way $Q + 1$ populations of accepted particles are generated. This first of these populations is simply a random sample from the prior distribution, with each subsequent population gradually evolving towards the target posterior distribution, $\pi(\eta|d(y^*, y_0) < \epsilon_Q)$.

The choice of predefined inputs can greatly impact the efficiency of the ABC-SMC algorithm. The number of populations to sample, $Q + 1$, and the tolerance vector, ϵ, are usually chosen so that the intermediate populations gradually evolve until they represent a sample from the target posterior, $\pi(\eta|d(y^*, y_0) < \epsilon_Q)$. Small differences between successive tolerances, that is ϵ_i and ϵ_{i+1}, mean that intermediate populations look very similar, whereas large differences between successive tolerances can result in a low particle acceptance rate and a slow evolution towards the target posterior; we note that the special case where $Q + 1 = 1$ corresponds to the ABC rejection algorithm, which in general performs poorly compared to

the ABC-SMC algorithm with $Q+1 > 1$ [127, 128]. Often the number of populations and corresponding tolerance vector are tuned by hand using trial and error [27, 127], however algorithms have been proposed to select these inputs in a more optimal manner. Rather than initialising the tolerance vector at the first step in the above algorithm, one approach is an adaptive choice of threshold schedule that consists of selecting the p-th quantile of distances between the simulated and observed data [22, 25, 64, 66]. Selecting the tolerance vector adaptively often improves the efficiency of the ABC-SMC algorithm [27], however Silk et al. [99] recommend caution as in some cases the algorithm may not converge to the posterior distribution for some values of p. In this work we choose to employ an adaptive approach to selecting the tolerance vector, the details of which are given in Appendix E and discussed in the next section.

The choice of perturbation kernel requires finding a balance between effectively exploring parameter space and the speed of convergence to the target posterior; a local perturbation kernel produces particles close to those from the previous population that will have a high probability of being accepted, provided ϵ is suitably chosen, whereas a widely spread kernel enables a fuller exploration of parameter space at the cost of a lower acceptance rate [27]. Thus, the choice of perturbation kernel also greatly impacts the efficiency of the ABC-SMC algorithm, the construction of which remains an unsolved problem in the context of ABC [27].

A variety of perturbation kernels have been proposed for use in an ABC-SMC context that range in their algorithmic complexity, from simpler component-wise perturbation kernels [27, 127, 128] to more the complex multivariate kernels [27, 66]. Filippi et al. [27] observed that component-wise kernels performed poorly in cases where parameters are highly correlated, since they poorly reflect the structure of the true posterior distribution of the parameters. Multivariate kernels with a covariance matrix depending on the previous population of particles, however, were found to be superior producing a higher acceptance rate, with a multivariate normal kernel with optimal local covariance matrix performing best [27].

It is important to consider, however, the increased algorithmic complexity that these more complex kernels bring alongside their increased particle acceptance rate. Filippi et al. [27] note that the computational cost of simulating the data for each particle often outweighs the complexity added to the ABC-SMC algorithm by a more complex kernel, although in cases where a simpler kernel produces the same acceptance rate, then the kernel that has a cheaper algorithmic

complexity should be chosen.

In this work, we employ a component-wise uniform perturbation kernel for simplicity. This consists of perturbing each component $1 \leq j \leq k$ of the parameter vector $\eta = (\eta_1, ..., \eta_k)$ independently according to a uniform distribution with width $2\sigma_j$. These widths of the uniform distributions can be fixed or a commonly chosen approach is to adaptively chose their value informed by the range of values for each component in the previous population [27]. We follow the latter approach and, therefore, index the kernel and its width by the population index, q. Thus, we define the perturbation kernel for population q (for $q = 1, ..., Q$) as

$$K_q(\eta|\eta^*) = \{\eta_j^* + U(-\sigma_{q,j}, \sigma_{q,j})\}_{j=1,...,k}, \tag{4.5}$$

where $\sigma_{q,j}$ is given by,

$$\sigma_{q,j} = \delta(\max\{\eta_{q-1,j}^{(1)}, ..., \eta_{q-1,j}^{(R)}\} - \min\{\eta_{q-1,j}^{(1)}, ..., \eta_{q-1,j}^{(R)}\}). \tag{4.6}$$

We note that the notation $\eta_{q-1,j}^{(i)}$ denotes the jth component of the ith particle of population $q-1$ and η_j^* is the jth component of η^*, a weighted sample from the previous population of accepted particles, $\{\eta_{q-1}^{(1)}, ..., \eta_{q-1}^{(R)}\}$. Further, the parameter $\delta \in \mathbb{R}$ determines the width of the perturbation kernel; a larger value of δ results in a larger region of parameter space being explored at each population iteration, which, again, comes with the potential cost of a lower particle acceptance rate.

4.4.2 Inferring the parameters of the gamma distributions for N_E and N_P

As discussed in Section 4.3, we assume the parameters N_E and N_P - the sizes of the N population of cells when the EGFR and PDGFR amplified populations of cells are introduced into the growing tumour, respectively - are gamma distributed as follows:

$$N_E \sim N_I + \Gamma(k_E, \theta_E) \quad \text{and} \quad N_P \sim N_I + \Gamma(k_P, \theta_P),$$

where $k_E, k_P > 0$ are the shape and $\theta, \theta_P > 0$ are the scale parameters and N_I is the size of the initial population of N cells, as defined in Chapter 3. In this section, we briefly explore how the choice of shape and scale parameters impact the spread of our simulated GBM cohorts and employ the ABC-SMC algorithm to test if we can infer these parameters for a set of synthetic data when all other

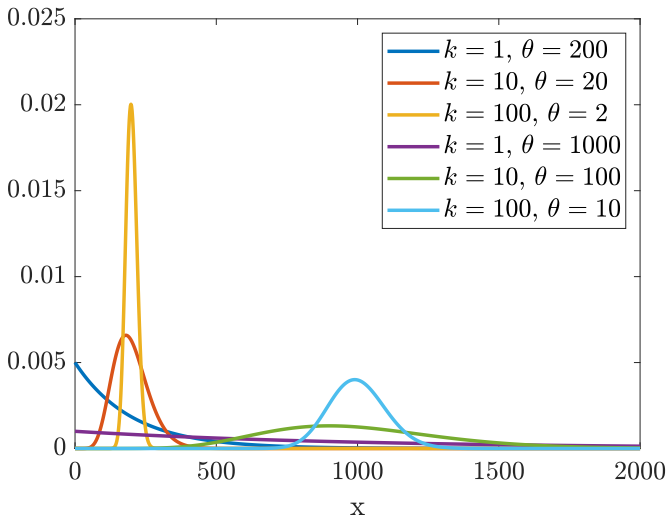


Figure 4.5: Probability density function for the gamma distribution with different values for the shape parameter, k, and scale parameter, θ.

parameters are known.

The shape and scale parameters, k and θ, of a gamma distribution, despite their names, both affect the shape of the gamma distribution, while its mean is given by $k\theta$. Figure 4.5 shows the probability density function (pdf) of the gamma distribution for six different pairs of k and θ. Of these distributions, three have a mean equal to 200 and three a mean of 1000 with the shape parameter k taking the values 1, 10 and 100 and θ the corresponding values in each case. From this, we see that the two pdfs with $k = 1$ and high values of θ are both curves monotonically decreasing from $x = 0$, with the higher value of θ having a lower maximum at $x = 0$. When $k = 10$ or 100, the pdf has a peak at the mean, with a higher value of k and lower value of θ resulting in a taller and narrower peak. Thus, the choice of k and θ will affect the spread of the N_E and N_P values sampled from each of the gamma distributions as well as the mean.

To explore how the choice of shape and scale parameters effects our simulated GBM cohorts, we produce a simulated GBM cohort of 100 GBMs, as described in Section 4.3.1, for each of these six (k, θ) pairs. In each of these simulated cohorts, the shape and scale parameters of the gamma distributions for N_E and N_P are equal, that is $k_E = k_P = k$ and $\theta_E = \theta_P = \theta$, taking the value pairs $(k, \theta) = (1, 200), (10, 20), (100, 2), (1, 1000), (10, 100)$ and $(100, 10)$. Meanwhile, all other model parameters are kept the same for each of the simulated cohorts and are as follows: $\nu^\rho = 1.3$, $\nu_P^\rho = 1.2$, $\nu_E^D = 1$, $\nu_P^D = 1$, $\alpha_{EP} = 0$ and $\alpha_{PE} = 0$. The results of these simulations are shown in Figures 4.6 and 4.7.

Figures 4.6(a)-(c) show the simulated cohorts with parameters $(k, \theta) = (1, 200)$, $(10, 20)$ and $(100, 2)$, respectively. Thus, in these figures the mean size of tumour

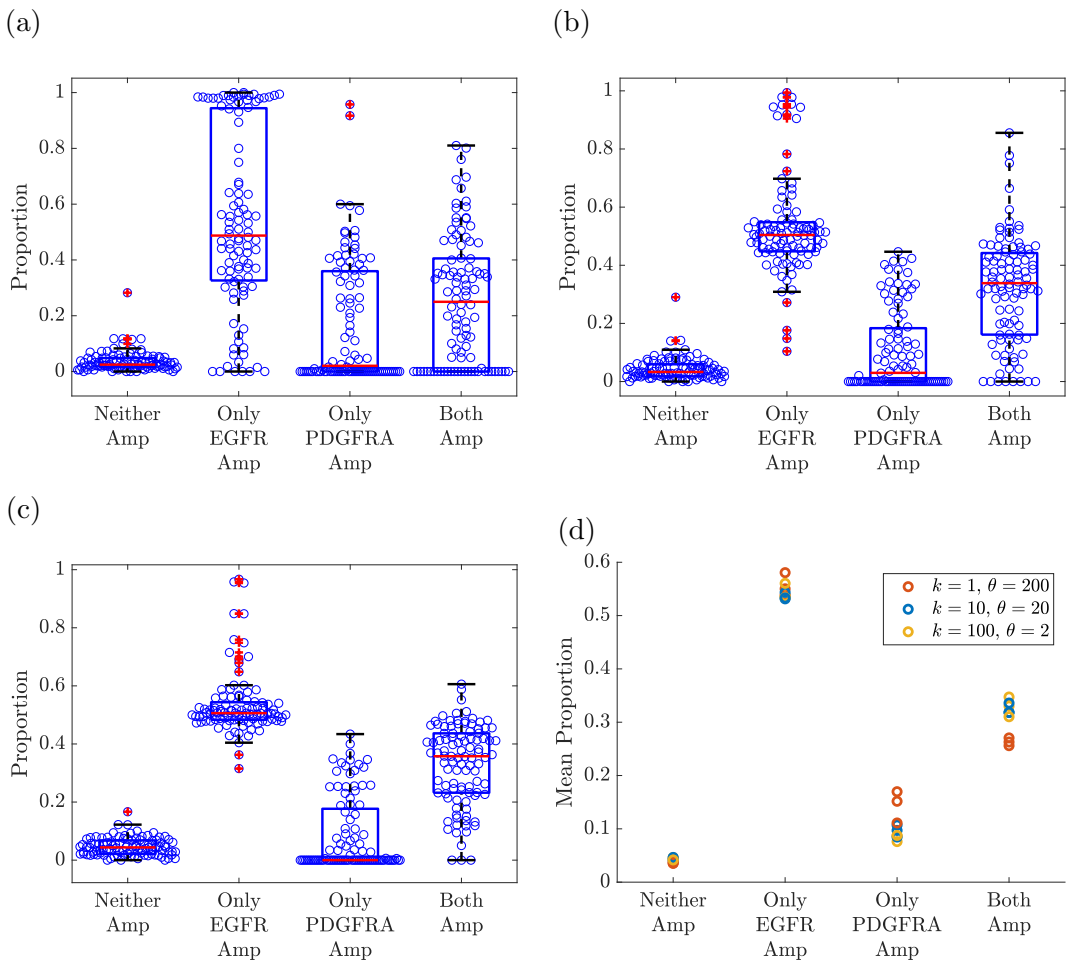


Figure 4.6: The results from simulated GBM cohorts with different parameters of the gamma distribution where the mean size of tumour when each amplified sub-population is introduced is kept fixed at 300 cells/mm^2. In each plot $k_E = k_P = k$ and $\theta_E = \theta_P = \theta$, where k and θ are varied. The box plots summarise the proportion of each simulated tumour that were determined to be amplified in neither gene (Neither), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both of the EGFR and PDGFRA genes (Both Amp) for the 100 simulations in each cohort of simulated GBMs with (a) $k = 1$, $\theta = 200$, (b) $k = 10$, $\theta = 20$, (c) $k = 100$, $\theta = 2$. A scatter plot showing the means of 3 simulated GBM cohorts for each (k, θ) pair is shown in (d). The other parameters used to simulate each GBM cohort were $\nu_E^\rho = 1.3$, $\nu_P^\rho = 1.2$, $\nu_E^D = 1$, $\nu_P^D = 1$, $\alpha_{EP} = 0$ and $\alpha_{PE} = 0$.

when each of the amplified sub-populations are introduced is 300cells/mm^2. The proportions of simulations with neither, only EGFR, only PDGFRA and both genes amplified for each of these simulated cohorts is shown as a boxplot for each case. These plots show the spread of data points across the cohort, with each blue circle representing the proportion of a single simulation that is or is not amplified, accordingly. From these boxplots, it is clear that the proportions of simulations with neither gene amplified is very similar for each of the simulated

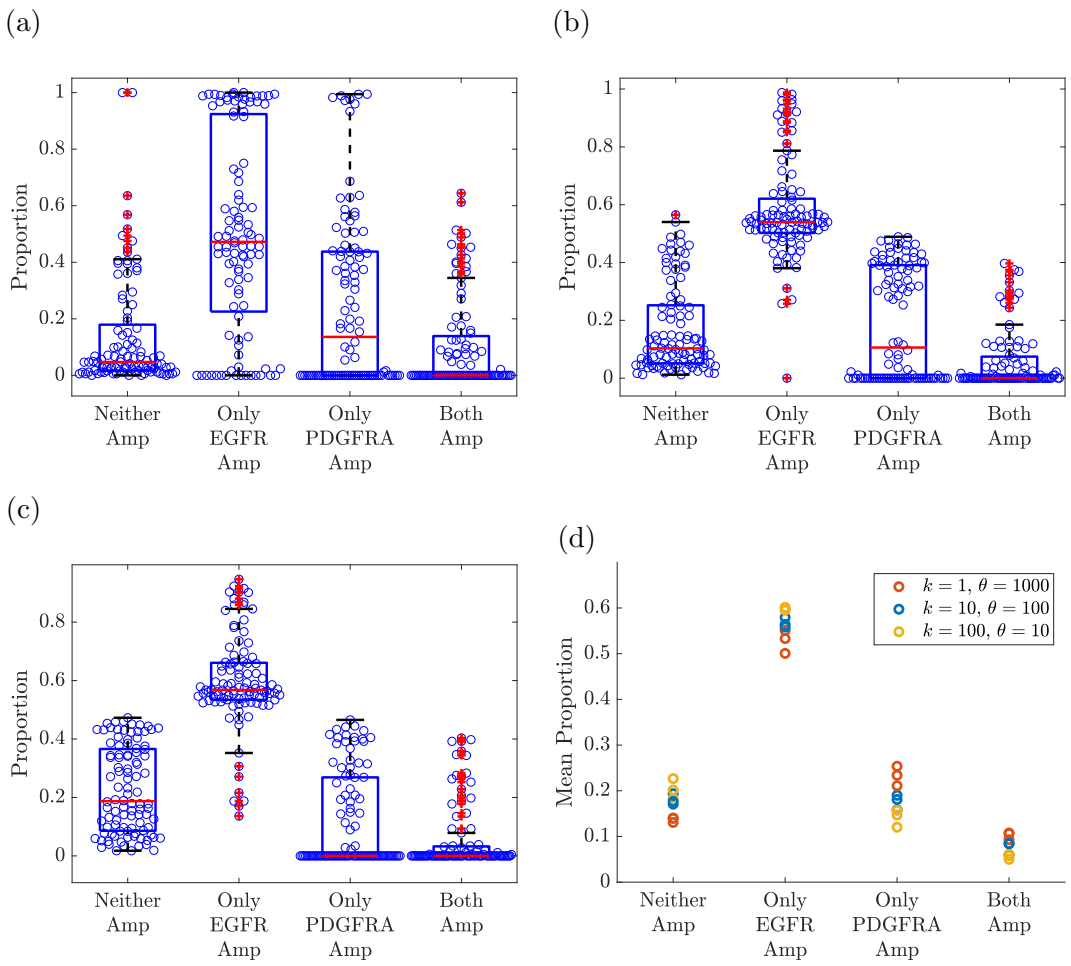


Figure 4.7: The results from simulated GBM cohorts with different parameters of the gamma distribution where the mean size of tumour when each amplified sub-population is introduced is kept fixed at 1100 cells/mm^2. In each plot $k_E = k_P = k$ and $\theta_E = \theta_P = \theta$, where k and θ are varied. The box plots summarise the proportion of each simulated tumour that were determined to be amplified in neither gene (Neither), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both of the EGFR and PDGFRA genes (Both Amp) for the 100 simulations in each cohort of simulated GBMs with (a) $k = 1$, $\theta = 1000$, (b) $k = 10$, $\theta = 100$, (c) $k = 100$, $\theta = 10$. A scatter plot showing the means of 3 simulated GBM cohorts for each (k, θ) pair is shown in (d). The other parameters used to simulate each GBM cohort were $\nu_E^\rho = 1.3$, $\nu_P^\rho = 1.2$, $\nu_E^D = 1$, $\nu_P^D = 1$, $\alpha_{EP} = 0$ and $\alpha_{PE} = 0$.

cohorts, each having all proportions low and tightly packed around the median, shown by the horizontal red line, with the exception of a few outliers. However, the spread of proportions that are amplified in one or both genes varies across the three simulated cohorts, with the biggest variation seen in the proportion of simulations that are amplified in only the EGFR gene. In Fig. 4.6(a) we can see that the simulated cohort contains some simulations that are amplified in the EGFR gene everywhere and others that are amplified nowhere, which mirrors

the range of amplification levels that we see in the patient cohort in Fig. 4.4a. As the shape parameter k increases, the spread of proportions that are EGFR amplified decreases and the proportions become more closely distributed around the median, shown by the red horizontal line on the boxplots. This is intuitively as expected, since the shape of the gamma distribution for these three sets of parameters varies considerably, as shown in Fig. 4.5, and the sampled values of N_E and N_P will be more spread out when $k = 1$ and the gamma pdf is very flat than when $k = 100$ and the pdf is a tall, narrow peak. Figures 4.7(a)-(c) show analogous results for three simulated cohorts of 100 GBMs with the parameters of the gamma distributions equal to $(k, \theta) = (1, 1000), (10, 100)$ and $(100, 10)$, respectively. A similar pattern to the spread of amplified and non-amplified proportions of simulations is observed, with the simulated cohort where $k = 1$ producing a wider range amplification pattern, more closely resembling that observed in the patient cohort in Fig. 4.4a.

Figures 4.6d and 4.7d show the mean proportions of simulations that are amplified and non-amplified in each simulated cohort. For each of the six (k, θ) pairs, three simulated cohorts are produced and the mean of each is calculated. These figures illustrate that the mean proportions for each simulated cohort can vary quite significantly when all model parameters are kept fixed. This is due to the stochasticity introduced to the model simulations as described in Section 4.3. While this stochasticity is more representative of the natural variation observed in a biological population, it may have implications for parameter inference; this will be explored in the remaining work in this chapter.

Despite the stochasticity of the results in Figures 4.6d and 4.7d, it is clear that the mean of the gamma distributions for N_E and N_P plays an important role in the mean amplification patterns observed for a simulated cohort of GBMs. In Fig. 4.6d, the amplified populations are introduced when the growing tumour is at a mean size of 300 cells/mm^3 leading to a larger mean proportion of amplified tumours than when introduced at a larger mean tumour size of 1100 cells/mm^3 in Fig. 4.7d. Thus, the products $k_E\theta_E$ and $k_P\theta_P$ will also be important to consider when inferring the parameters as well as their individual values. This was expected to be the case, since we found that the model was highly sensitive to the parameters N_E and N_P in the sensitivity analysis presented in the previous chapter and, thus, we expect that the model would also be sensitive to their mean values now that we are assuming they are gamma distributed.

The inferability of model parameters is linked to the sensitivity of the model output to those parameters; indeed, if varying a particular model parameter has

very little impact on the output of the model, then it will be difficult to infer that parameter [127]. As we found in the previous chapter that the amplification patterns observed in the model were very sensitive to the parameters N_E and N_P, it is intuitive to expect that the mean amplification patterns observed when using the model to generate a cohort of GBMs will be sensitive to the means of the gamma distributions, $k_E\theta_E$ and $k_P\theta_P$, for N_E and N_P and that the means will be inferable. However, it may be the case that using the mean amplification patterns alone may not be enough information to be able to infer the parameters k_E, θ_E, k_P and θ_P individually. For this reason, we choose to use the standard deviation of the amplified proportions to infer the parameters in our model to as well. As shown in Fig.s 4.6 and 4.7, the spread of proportions of non-amplified and amplified tumours seen in a simulated cohort of GBMs varies as k_E, θ_E, k_P and θ_P are varied and so we expect that incorporating the standard deviation into our parameter inference will help to identify these parameters. Thus, we use eight data points from the dataset we are inferring the model parameters for, which we denote as two vectors y_0^M and y_0^{SD}. The vector y_0^M contains the four mean values of the tumour proportions amplified in neither, only EGFR, only PDGFRA and both genes in the dataset. The vector y_0^{SD} contains the four standard deviations of these four sets of proportions. For each particle sampled in the ABC-SMC algorithm, we generate a simulated cohort of GBMs representative of the size of the dataset we are inferring the parameters for and calculate the analogous eight data points, calculating the proportions of tumours amplified as previously described. We then denote these eight points as the vectors y_M^* and y_{SD}^*, each consisting of the four mean and four standard deviation values, respectively. We use the Euclidean distance as the distance metric, d, in the ABC-SMC algorithm and choose to calculate the distance between y_0^M and y_M^* separately to the distance between y_0^{SD} and y_{SD}^*. In this way, the tolerance vector is defined as $\epsilon = (\epsilon_1, ..., \epsilon_Q)$, where $\epsilon_i = (\epsilon_i^M, \epsilon_i^{SD})$ for $i = 1, ..., Q$. As mentioned before, the tolerance vector is chosen adaptively during the ABC-SMC algorithm, the process used for determining it is detailed in Appendix E. A particle is then accepted into population i in the ABC-SMC algorithm (line 19 of the algorithm in Section 4.4.1) if

$$d(y_M^*, y_0^M) < \epsilon_i^M \quad \text{and} \quad d(y_{SD}^*, y_0^{SD}) < \epsilon_i^{SD}. \tag{4.7}$$

Accepting the particles in this way allows us to infer parameters with a desired tolerance level to both the y_0^M and y_0^{SD} and avoid a situation where we may end

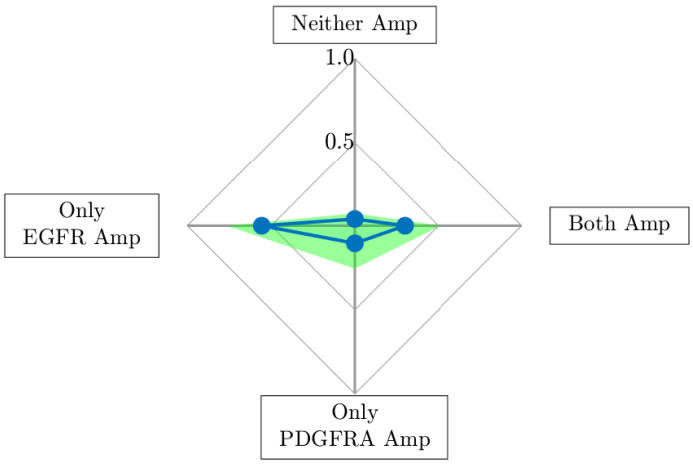


Figure 4.8: A spider plot showing the mean proportions of simulations in a simulated cohort of 100 GBMs that were amplified in neither gene, only the EGFR gene, only the PDGFRA gene and both gene; these four data points are represented in blue. The green shaded area shows the mean ± the standard deviation on each axis. The simulated cohort was produced as described in Section 4.3.1 with model parameters $(\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP}) = (1.3, 1.2, 1, 1, 10, 20, 10, 20, 0, 0)$.

up fitting to the mean data points well and the standard deviation points poorly, for example.

Inferring the model parameters $k_E = k_P = 10$ and $\theta_E = \theta_P = 20$

In order to now test whether we can infer the shape and scale parameters of the gamma distributions for N_E and N_P in our model, namely parameters k_E, k_P, θ_E and θ_P, we generate a synthetic dataset and test if we are able to recover these parameters, while assuming that all other model parameters are known. The synthetic dataset is generated by producing a simulated cohort of 100 GBMs, as described in Section 4.3.1, with parameters

$$(\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP}) = (1.3, 1.2, 1, 1, 10, 20, 10, 20, 0, 0).$$

The proportion of each simulation that is amplified in neither gene, only EGFR, only PDGFRA and both genes is calculated as previously described. The mean and standard deviation of these four measures for the 100 simulated tumours in the synthetic dataset are shown in Figure 4.8. From this spider plot, we see that the mean of the proportions of simulated tumours with only the EGFR amplified is more than half at 0.5556, while the mean proportions with both genes, only PDGFRA and neither gene are 0.2996, 0.1046 and close to zero at 0.0403,

respectively. This mean amplification pattern was to be expected, since both the EGFR and PDGFRA amplified populations were introduced into the growing tumour early, at a mean size of 300cells/mm^2, and both had a proliferative advantage over the non-amplified tumour cell population, with the EGFR amplified population having the greater advantage. Thus, we would expect the proportion of tumours with only EGFR amplified to be highest, followed by the PDGFRA and both genes amplified proportions since the interactions between these populations were neutral. Finally, we would expect the non-amplified proportion of simulated tumours to be small.

The results seen here are similar to those shown in Fig. 4.6b, which shows another simulated cohort produced with the same parameters. While the cohorts are similar, the results differ slightly each time due to the stochastic nature of the model and its parameters. Indeed, three additional runs with the same parameter set produced simulated cohorts with distances from the means and standard deviations of

$$(0.0405, 0.0274, 0.0576) \quad \text{and} \quad (0.0104, 0.0345, 0.0216), \tag{4.8}$$

respectively. Thus, when implementing the ABC-SMC algorithm we should not expect it to return a final population of accepted particles with distances from the mean and standard deviations less than these values, although achieving final tolerances, ϵ_Q^M and ϵ_Q^{SD}, close to these distances is desired.

We now implement the ABC-SMC algorithm as described in Section 4.4.1 with $Q+1 = 8$ populations of $R = 200$ sampled particles, to infer the unknown parameter vector $(k_E, \theta_E, k_P, \theta_P)$ for the simulated cohort shown in Fig. 4.8. As the synthetic dataset consists of 100 GBMs, for each particle sample a simulated cohort of 100 GBMs is produced. The prior distributions for the unknown parameters are defined such that each of the parameter pairs (k_E, θ_E) and (k_P, θ_P) are jointly uniformly distributed over the region S_i, where S_i is defined as

$$S_i = \{k_i, \theta_i \in \mathbb{R}^2 : 0.001 \leq k_i \leq 200,\ 0.001 \leq \theta_i \leq 1200,\ k_i\theta_i < 3000\}, \tag{4.9}$$

for $i = \{E, P\}$. These prior distributions are defined in this way so that feasible parameter ranges are sampled from without specifying any preference to particular parameter values, except to exclude those values giving means of the gamma distribution from which N_E and N_P are subsequently sampled larger than 3000. The shape and scale parameters of the gamma distributions are chosen to have joint prior distributions, since it is known that the model output greatly depends

on their product, that is the mean of the gamma distributions, as well as their individual values. While it is known that the amplified populations are not introduced at such late mean times, defining the regions S_i in this way greatly reduces the size of the parameter space from which the initial population of particles are sampled and will improve the performance of the inference algorithm.

As described in Appendix E, the tolerance vectors are generated using an adaptive approach, yielding

$$\epsilon^M = (0.4640, 0.3879, 0.3173, 0.2432, 0.1794, 0.1364, 0.1076)$$

and

$$\epsilon^{SD} = (0.1998, 0.1752, 0.1436, 0.1018, 0.0773, 0.0591, 0.0471)$$

in this instance. Therefore, the final population of accepted particles satisfies the conditions,

$$d(y^*_M, y^M_0) < 0.1076 \quad \text{and} \quad d(y^*_{SD}, y^{SD}_0) < 0.0471, \tag{4.10}$$

where y^*_M and y^*_{SD} denote the four mean and four standard deviation data points of a simulated cohort generated with a given parameter particle and y^M_0 and y^{SD}_0 denote the same eight data points of the synthetic data. Upon comparing these final values to the distances from the synthetic data observed in (4.8) following multiple runs with the true parameters, we find that the ABC-SMC algorithm has produced a final population of accepted particles with distances close to these values.

The results of this implementation of the ABC-SMC algorithm are summarised in Fig. 4.9. The scatter plots show the accepted particles in the first (blue), third (red), sixth (yellow) and final/eighth (purple) populations plotted in (a) k_E–θ_E, (c) k_P–θ_P and (e) $k_E\theta_E$–$k_P\theta_P$ parameter spaces. These scatter plots illustrate how the populations of accepted particles gradually evolve from a random sample from the prior distribution towards a final population representing a sample from the target posterior distribution, $\pi(\eta|d(y^*, y_0) < \epsilon_Q)$. Each subsequent population of accepted particles occupies a smaller region of the parameter space as the cloud of accepted particles converge towards the final population. The second, fourth and seventh populations follow the same pattern and are only omitted here to avoid the scatter plots becoming too crowded.

Plots (b), (d) and (f) show kernel density estimates of the parameter values in the final population of accepted particles, from which we can see that the ABC-

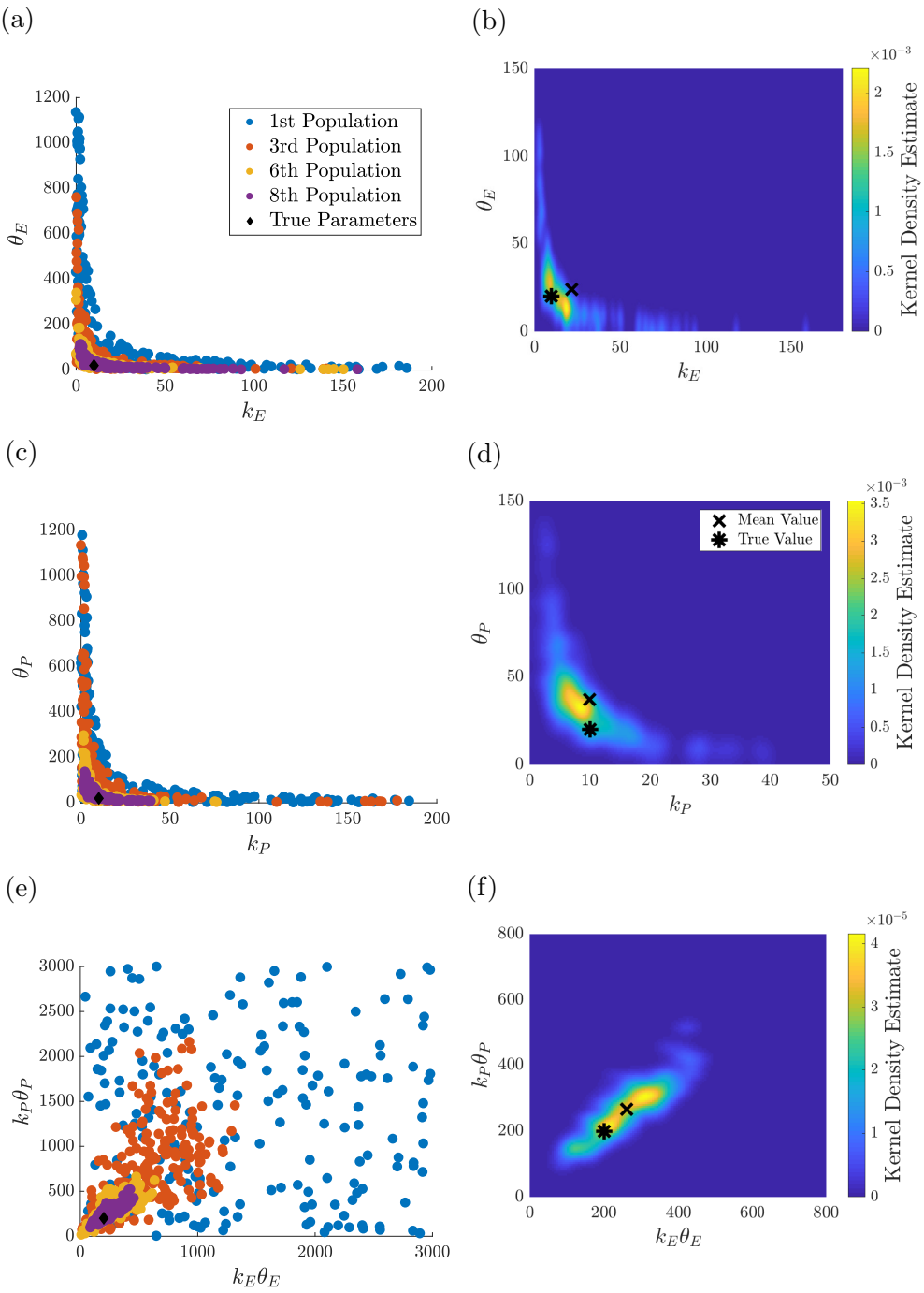


Figure 4.9: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters k_E, k_P, θ_E and θ_P, assuming all other parameters are known. Scatter plots (a), (c) and (e) show the initial (blue), third (red), sixth (yellow) and final (purple) populations of accepted particles for parameters (a) k_E and θ_E; and (c) k_P and θ_P; (e) $k_E\theta_E$ and $k_P\theta_P$. The true parameter values $k_E = k_P = 10$ and $\theta_E = \theta_P = 20$ used to create the synthetic data are shown by a black diamond. Plots (b), (d) and (f) show kernel density estimates of the final population of accepted particles from the ABC-SMC algorithm for parameters (b) k_E and θ_E; (d) k_P and θ_P; and (f) $k_E\theta_E$ and $k_P\theta_P$.

SMC algorithm has performed fairly well. Figure 4.9b shows that the region of high density of accepted parameter values of k_E and θ_E include the true parameter values. The distribution for the k_E parameter does extend the right of this peak, however, illustrating that a number of parameter particles were accepted into the final population with much higher values of k_E, with the mean value of the parameter k_E in the final population of accepted particles being 21.85, approximately double the true value. Meanwhile, the mean value of the θ_E parameter is 24.06. The kernel density estimates of the parameters k_P and θ_P in the final population of particles are shown in Fig. 4.9d. From this we see that the accepted parameter values are distributed closely around the true values, with the region of highest density lying slightly above the true parameter values. The mean value of k_P in the final population is 9.89, which is very close to the true value of 10, while the mean value of θ_P is 36.92. Finally, Fig. 4.9f shows the density distribution of accepted values for the means of the gamma distributions, $k_E\theta_E$ and $k_P\theta_P$. From this, we can see that the accepted mean values have converged well around the true value of 200, with the mean values of $k_E\theta_E$ and $k_P\theta_P$ in the population of particles being 260.87 and 265.26, respectively. These are slightly larger than the true values and we also see that the region of highest density mostly includes slightly higher parameter values as well. This may be by chance due to the random sample of parameters, or it may be the result of the stochastic nature of the model meaning that the synthetic data produced fit these slightly higher mean values better; sampling more particles and fitting to the mean data points from multiple sets of synthetic data would be likely to help to resolve this, although this may not be possible when using real sets of data.

Depicting the results in this way gives an insight into the relationship between model parameters. From Fig.s 4.9(e) and (f), we observe that the products $k_E\theta_E$ and $k_P\theta_P$ of the accepted parameter particles cluster around the straight line $k_E\theta_E = k_P\theta_P$ for values between approximately 50 and 500. This indicates that accepted particles in the final population are those with approximately equal values for the mean of the gamma distributions for the N_E and N_P parameters. As we know that the true values used to generate the synthetic data were indeed equal, this indicates that the ABC-SMC algorithm has performed well and identified a target posterior for the means of the gamma distributions that is consistent with the true, known values. Meanwhile, Fig.s 4.9(a) and (b) show that the accepted parameter values for k_E and θ_E are highly correlated, converging around the line $k_E\theta_E \approxeq 200$, with a similar relationship between parameters k_P and θ_P shown in Fig.s 4.9(c) and (d). This, again, illustrates that the inference

algorithm is able to infer the mean of gamma distributions well.

From the scatter plots in Fig. 4.9(a) and (c), we observe that we are able to infer the individual parameters to some degree. Toni [127] considers the inferability of parameters using the ABC-SMC algorithm in the following sense: *if a posterior distribution is considerably narrower than the prior distribution, then the corresponding parameter is inferable, while if the posterior and prior distributions are similar, then the parameter is not inferable.* Upon comparing the final population (purple points) to the initial random sample from the prior distribution (blue points), we can see that they are quite different; the inference algorithm has excluded large values for the parameters k_E, θ_E, k_P and θ_P and the posterior distributions are narrower than the prior distribution.

Inferring the model parameters $k_E = k_P = 1$ and $\theta_E = \theta_P = 1000$

We now repeat the same analysis to test whether we can infer the k_E, θ_E, k_P and θ_P parameters for synthetic data generated with different values of these parameters, namely $(k_E, \theta_E, k_P, \theta_P) = (1, 1000, 1, 1000)$, while all other model parameters are kept the same as for the previous set of synthetic data. This new set of synthetic data is produced in the same way as before by creating a simulated population of 100 GBMs and is shown in Fig. 4.10. Here we see that introducing the amplified sub-populations at later times, when the tumour is on average a larger size at 1100 cells/mm^2, produces a slightly different mean amplification pattern. The EGFR and PDGFRA populations are still able to grow so that the mean proportions of tumours amplified in these genes are still quite high because of their proliferative advantage over the non-amplified population of cells, however this later mean introduction time has resulted in a slightly larger non-amplified mean proportion.

Notably different between the two synthetic datasets is the size of the green area on the spider plots. As N_E and N_P are gamma distributed with scale parameters $k_E = k_P = 1$ when generating this synthetic data, a wide variety of introduction times for the amplified sub-populations will be selected as their gamma distributions will be very flat, as shown in Fig. 4.5. Thus, we see a wider variety of amplified and, consequently, non-amplified proportions across the simulated cohort of GBMs resulting in larger standard deviations. We also see the impact of this wider variety when comparing three additional datasets generated with the same parameters to the set of synthetic data shown in Fig. 4.10, where we observe distances from the means, $d(y^*_M, y^M_0)$, and standard deviations, $d(y^*_{SD}, y^{SD}_0)$, of $(0.0603, 0.0936, 0.1038)$ and $(0.0487, 0.0835, 0.1005)$, respectively.

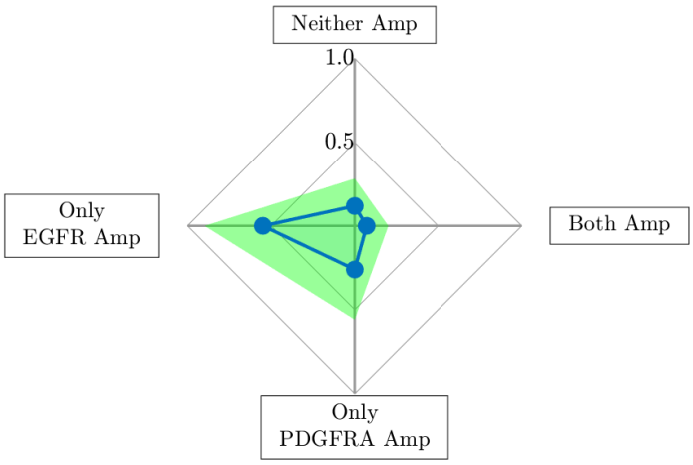


Figure 4.10: A spider plot showing the mean proportions of simulations in a simulated cohort of 100 GBMs that were amplified in neither gene, only the EGFR gene, only the PDGFRA gene and both gene; these four data points are represented in blue. The green shaded area shows the mean + the standard deviation on each axis. The simulated cohort was produced as described in Section 4.3.1 with model parameters $(\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP}) = (1.3, 1.2, 1, 1, 1, 1000, 1, 1000, 0, 0)$.

As these distances are larger than for the previous set of synthetic data, we do not expect the ABC-SMC algorithm to achieve tolerance levels as low as previously, however somewhere close to these values is desired.

We now employ the ABC-SMC algorithm to infer the $(k_E, \theta_E, k_P, \theta_P)$ parameters for the synthetic dataset shown in Fig. 4.10. All details of the ABC-SMC algorithm are kept the same as before, including the definition of the prior distributions. The tolerance vectors generated adaptively as the algorithm progresses are

$$\epsilon^M = (0.3950, 0.2971, 0.2391, 0.1943, 0.1565, 0.1325, 0.1124)$$

and

$$\epsilon^{SD} = (0.2323, 0.2076, 0.1822, 0.1654, 0.1400, 0.1197, 0.1036).$$

The final tolerance values achieved are good when taking into account that a second dataset produced with the true parameters gave distances from the means and standard deviations of the synthetic dataset of 0.1038 and 0.1005, respectively.

The results of this inference are shown in Fig. 4.11. We are able to infer some information about the k_E and θ_E parameters from these results, as large values of k_E and small values of θ_E are excluded from the final population of accepted particles in Fig.s 4.11(a) and (b). The majority of particles are spread

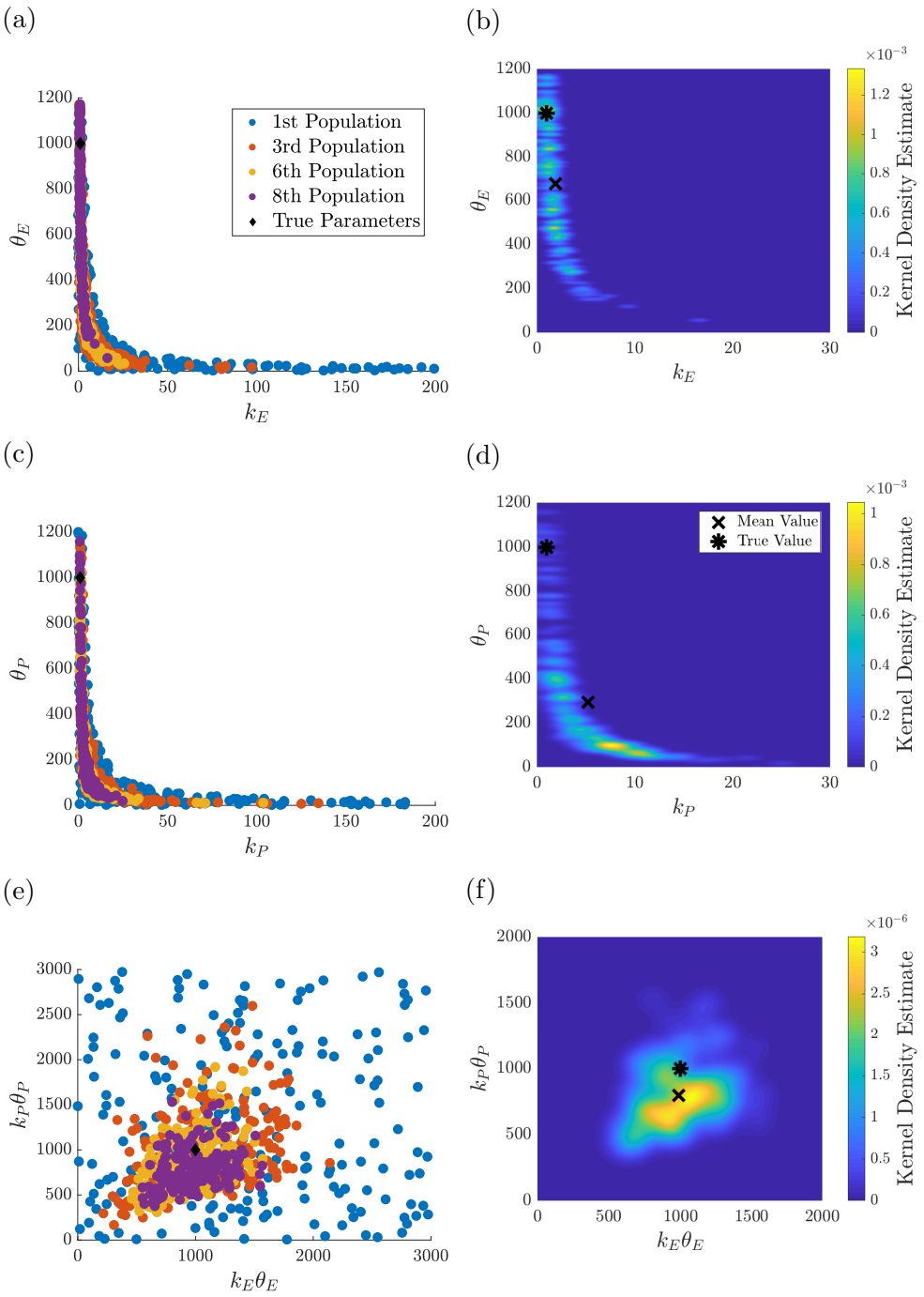


Figure 4.11: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters k_E, k_P, θ_E and θ_P, assuming all other parameters are known. Scatter plots (a), (c) and (e) show the initial (blue), third (red), sixth (yellow) and final (purple) populations of accepted particles for parameters (a) k_E and θ_E; (c) k_P and θ_P; and (e) $k_E\theta_E$ and $k_P\theta_P$;. The true parameter values, $k_E = k_P = 1$ and $\theta_E = \theta_P = 1000$, used to create the synthetic data are shown by a black diamond. Plots (b), (d) and (f) show kernel density estimates of the final population of accepted particles for parameters (b) k_E and θ_E; (d) k_P and θ_P; (f) $k_E\theta_E$ and $k_P\theta_P$.

in a fairly even density in a band that includes the true parameter values, with θ_E values ranging from approximately 200 to 1200 and no particular regions of higher density forming. The mean parameter values in the final population of accepted particles are 1.90 and 675.81 for k_E and θ_E, respectively.

From Fig. 4.11(c) and (d), we see that the k_P and θ_P parameters have not been inferred as well, since we see a region of higher density forming at lower values of θ_P with k_P between approximately 8 and 10, far away from the true parameter values. The mean parameter values for k_P and θ_P in the final population of accepted particles are 5.20 and 293.55, respectively. We expect these parameters have been inferred less well than the k_E and θ_E parameters due to chance, either in the particles sampled or in the generation of the synthetic data.

Finally, from Fig.s 4.11(e) and (f), we can see that the algorithm has been able to infer the means of the gamma distributions for N_E and N_P, that is $k_E\theta_E$ and $k_P\theta_P$. We can see that a region of high density has formed close to the true parameter values, with mean values of $k_E\theta_E$ and $k_P\theta_P$ in the final population of accepted particles of 987.73 and 796.53, respectively. Thus, despite individual parameters not being inferred well, the final population of particles for the means, $k_E\theta_E$ and $k_P\theta_P$, has still clustered around the true parameters well, providing us with information about the means of the gamma distributions.

4.4.3 Inferring all model parameters using the ABC-SMC algorithm

In this section, we now test whether we can infer all model parameters using the ABC-SMC inference algorithm. Thus, instead of just inferring four parameters, we will now explore how well we can infer the parameters when all ten parameters, $(\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP})$, are unknown. We choose to implement the algorithm to infer the parameters for the synthetic dataset shown in Fig. 4.8, where the true parameters values used to generate the data were $(1.3, 1.2, 1, 1, 10, 20, 10, 20, 0, 0)$.

Thus, we now implement the ABC-SMC algorithm as described in Section 4.4.1 with $Q + 1 = 8$ populations of $R = 500$ sampled particles to infer the unknown parameter vector $(\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP})$. Once again, as the synthetic data is comprised of 100 GBMs, a simulated cohort of 100 GBMs is produced for each particle sampled throughout the inference. The prior distri-

butions for each unknown parameter are defined as follows:

$$\nu_E^\rho \sim U(1,2), \tag{4.11}$$
$$\nu_P^\rho \sim U(1,2), \tag{4.12}$$
$$\nu_E^D \sim U(1,2), \tag{4.13}$$
$$\nu_P^D \sim U(1,2), \tag{4.14}$$
$$\alpha_{EP} \sim U(-5,5), \tag{4.15}$$
$$\alpha_{PE} \sim U(-5,5), \tag{4.16}$$

and each of the parameter pairs (k_E, θ_E) and (k_P, θ_P) are jointly uniformly distributed over the region S_i, where S_i is defined as

$$S_i = \{k_i, \theta_i \in \mathbb{R}^2 : 0.001 \leq k_i \leq 200,\ 0.001 \leq \theta_i \leq 1200,\ k_i\theta_i < 3000\}, \tag{4.17}$$

for $i = \{E, P\}$. As previously, the tolerance vectors are generated adaptively as the inference algorithm progresses and, during this implementation, the values produced are

$$\epsilon^M = (0.5472, 0.4388, 0.3252, 0.2442, 0.1778, 0.1340, 0.1040)$$

and

$$\epsilon^{SD} = (0.2236, 0.1748, 0.1245, 0.0978, 0.0774, 0.0615, 0.0499).$$

The results of this parameter inference are shown in Fig.s 4.12–4.15. The first of these figures shows scatter plots of the parameter values in the first, third, sixth and eighth (final) populations of accepted particles, where we see the populations evolving gradually towards the final population, which gives us the final estimate of the posterior distribution from the inference. Fig.s 4.13–4.15 show plots of the estimated kernel density of the parameter values in the final population of accepted particles.

Figure 4.12: (See next page for figure.) Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters $(\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP})$ for the synthetic data with true parameters $(1.3, 1.2, 1, 1, 10, 20, 10, 20, 0, 0)$. Each scatter plot shows the first (blue), third (red), sixth (yellow) and final (purple) population of accepted particles plotted in (a) $k_E - \theta_E$, (b) $k_P - \theta_P$, (c) $\nu_E^\rho - \nu_P^\rho$, (d) $\nu_E^D - \nu_P^D$, (e) $\alpha_{EP} - \alpha_{PE}$, (f) $k_E\theta_E - k_P\theta_P$, (g) $k_E\theta_E - \nu_E^\rho$ and (h) $k_P\theta_P - \nu_P^\rho$ parameter space. The true parameter values are shown by the grey diamond on each scatter plot.

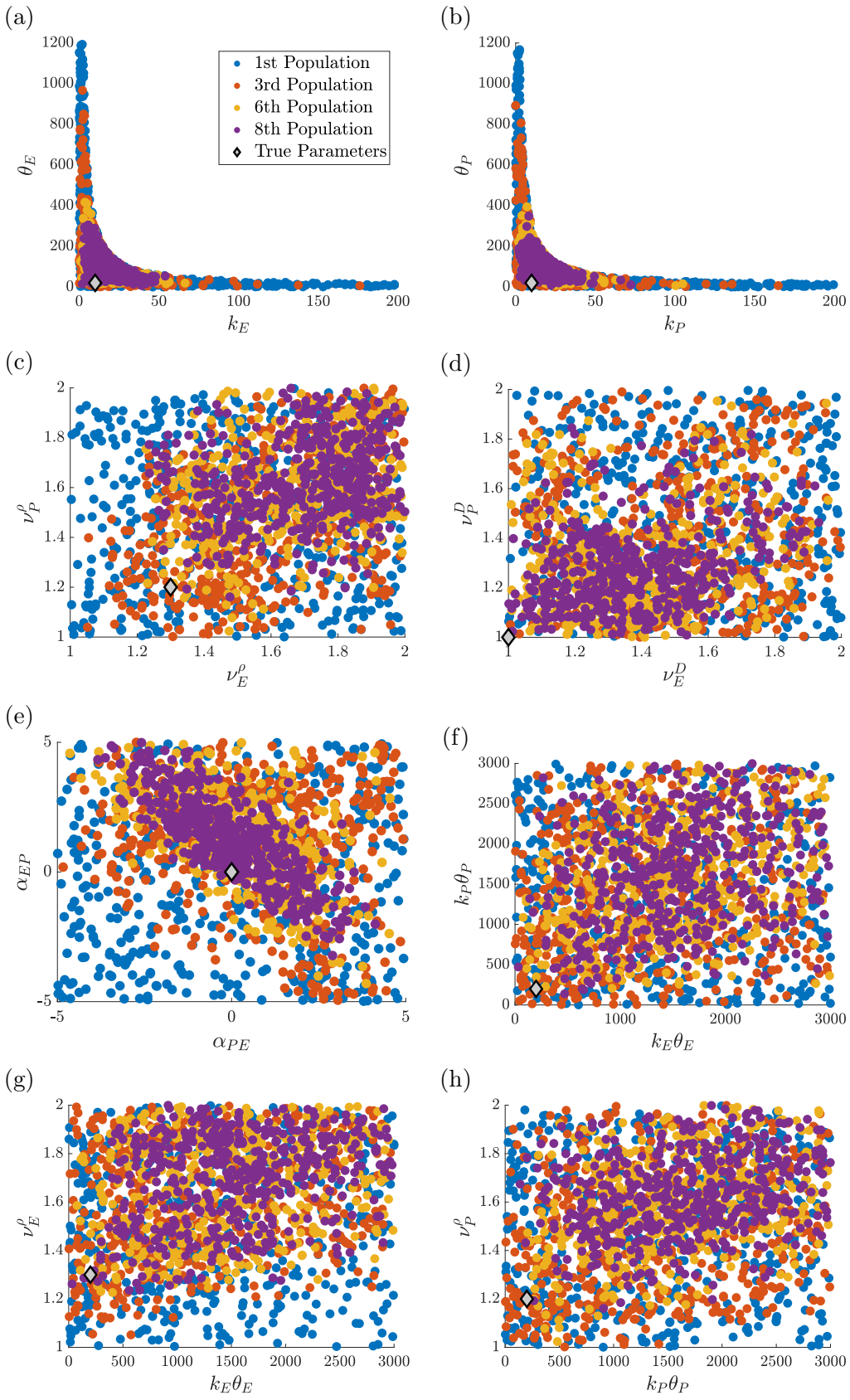

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
1st Population
3rd Population
6th Population
8th Population
True Parameters
θ_E
k_E
θ_P
k_P
ν_P^ρ
ν_E^ρ
ν_P^D
ν_E^D
α_{EP}
α_{PE}
$k_P\theta_P$
$k_E\theta_E$

Figure 4.12

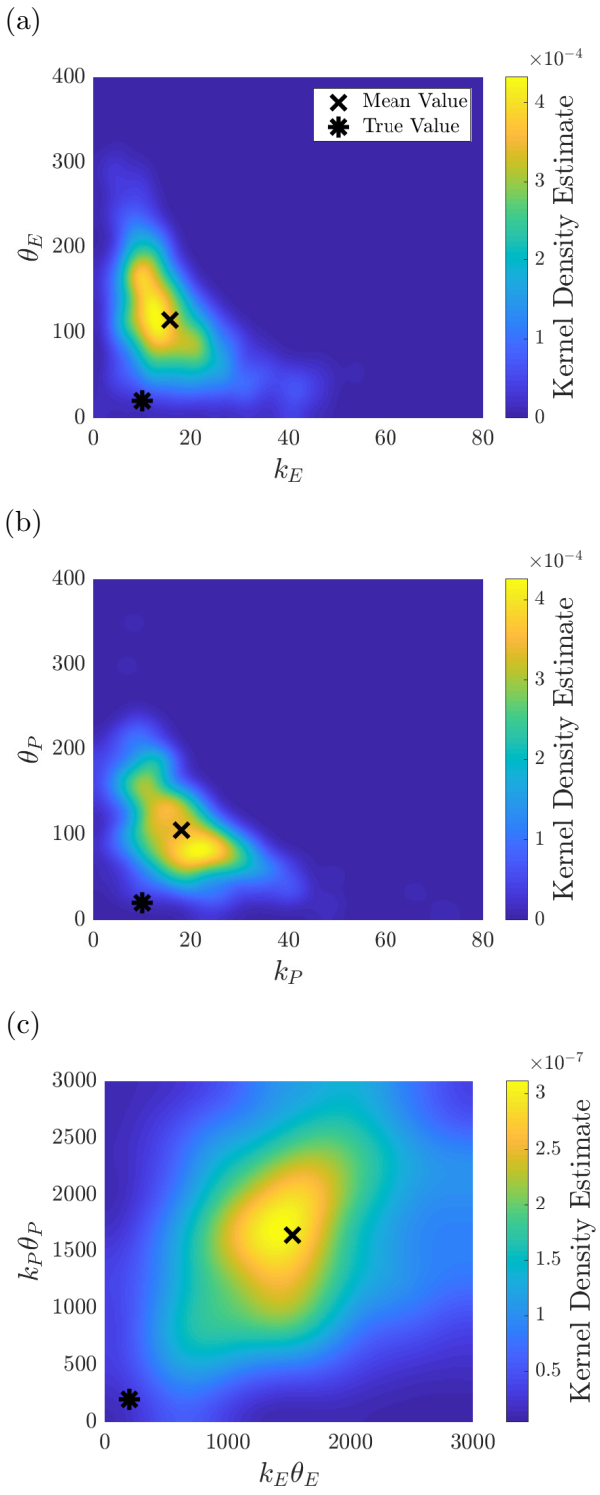


Figure 4.13: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters for the patient cohort data shown in Fig.4.8. Each plot shows kernel density estimates of the final population of accepted parameter particles in (a) $k_E - \theta_E$, (b) $k_P - \theta_P$ and (c) $k_E\theta_E - k_P\theta_P$ parameter space. The mean parameter values in the final population of accepted particles and the true parameter values are also shown on each plot.

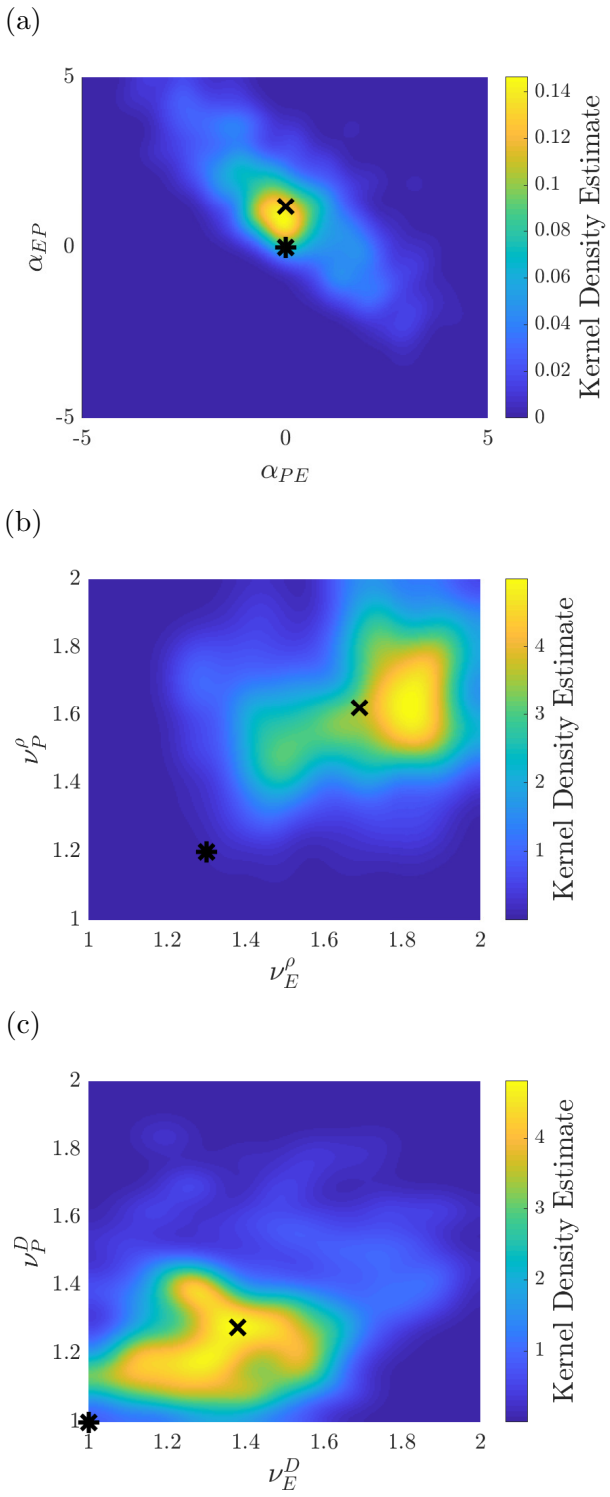


Figure 4.14: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters for the synthetic data shown in Fig.4.8. Each plot shows kernel density estimates of the final population of accepted parameter particles in (a) $\alpha_{PE} - \alpha_{EP}$, (b) $\nu_E^\rho - \nu_P^\rho$ and (c) $\nu_E^D - \nu_P^D$ parameter space. The mean parameter values in the final population of accepted particles and the true parameter values are also shown on each plot.

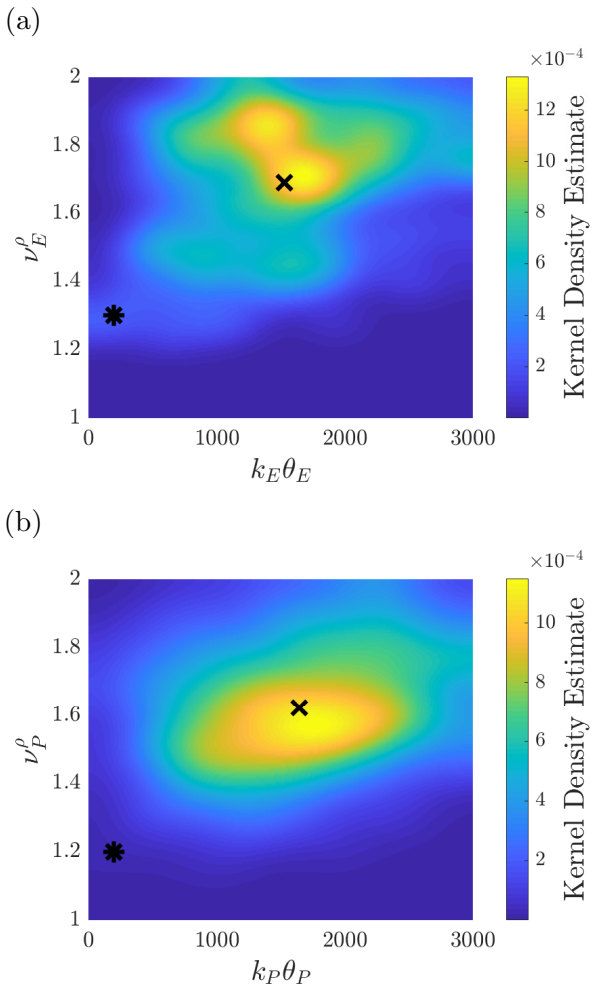


Figure 4.15: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters for the synthetic data shown in Fig.4.8. Each plot shows kernel density estimates of the final population of accepted parameter particles in (a) $k_E\theta_E - \nu_E^\rho$ and (b) $k_P\theta_P - \nu_P^\rho$ parameter space. The mean of parameter values in the final population of accepted particles and the true parameter values are also shown on each plot.

From plots (a) and (b) of both Fig.s 4.12 and 4.13 we can see that the algorithm has inferred some information about the parameters of the gamma distributions for N_E and N_P, that is k_E, θ_E, k_P and θ_P. The posterior distributions are narrower than the prior and the true parameter value is contained within the posterior distribution in each case, although slightly outside the regions of highest density in Fig.s 4.13(a) and (b). The parameters k_E and k_P have been inferred well, with mean values in the final population of 15.75 and 18.08, respectively. Meanwhile, we observe that θ_E and θ_P have not been inferred as well as in the previous parameter inference (see Fig. 4.9) when only these four parameters, $(k_E, \theta_E, k_P, \theta_P)$ were being inferred; this was to be expected due to the increased dimensionality of the parameter inference in this instance. Furthermore, we see

from Fig. 4.16(c) that the most likely values inferred for the means of the gamma distributions are much larger than the true parameter values, with the mean values for $k_E\theta_E$ and $k_P\theta_P$ in the final population of particles being 1527.7 and 1641.8, respectively. This, again, contrasts the results in Fig. 4.9, where the means of the gamma distributions for N_E and N_P were inferred well for the same set of synthetic data. If we look at the inference results for parameters ν_E^ρ and ν_P^ρ, a reason for this poor inference of the means of the gamma distributions for N_E and N_P becomes clearer. From Fig. 4.14(b), we see that the region of highest density of accepted particles in the final population is with values of ν_E^ρ between 1.7 and 1.9 and values of ν_P^ρ between 1.5 and 1.7; these are much higher than the true parameter values of 1.3 and 1.2, respectively. These higher proliferative advantages of the EGFR and PDGFRA amplified sub-populations mean that the populations can be introduced into the growing tumour at a later time and still grow to reach the same amplification levels as if they were introduced earlier. Thus, similar amplification patterns can be observed depending on the combination of proliferative advantage afforded to the amplified sub-populations and their time of introduction into the growing tumour. Indeed, as shown in Fig.s 4.15(a) and (b) a relationship between the values of the proliferative advantage and mean of the gamma distributions can be seen for each of the amplified sub-populations. Here we observe that no particles in the final accepted population have large vales of $k_E\theta_E$ and small values of ν_E^ρ, whereas particles with these later mean introduction times are accepted if the EGFR amplified sub-population is afforded a larger proliferative advantage; a similar pattern is observed for the parameters of the PDGFRA amplified sub-population. Therefore, due to the nature of the relationship between these model parameters, it is difficult to infer them well if both are unknown. However, some insight can still be gained about the parameters from this inference. The mean parameter values shown in Fig. 4.13(c) for $k_E\theta_E$ and $k_P\theta_P$ are both very similar, which is consistent with the knowledge that the true parameters used to generate the synthetic data were the same. Additionally, the posterior distributions for the proliferative advantages shown in Fig. 4.14(b) indicate that the EGFR amplified cells have a proliferative advantage over the PDGFRA amplified sub-population, which is also true of the populations in the synthetic data. Thus, while the majority of parameter values in the final population of accepted particles do not match the true parameter values well, the inference does still provide us with information about the EGFR and PDGFRA populations that is consistent with our knowledge about the synthetic data. Furthermore, it may also be the case that the parameter inference algorithm has

performed poorly due to the stochasticity in the model, as the particular set of synthetic data generated may simply fit these inferred parameter values better by chance. This is a challenge encountered when inferring parameters for models with stochasticity, as successive runs of the model with the same parameters can produce quite different outputs, as previously seen by the distances calculated between the synthetic data and three further runs of the model in (4.8). Fitting to a larger synthetic dataset or the mean of multiple datasets would be likely to help improve this. However, a single dataset consisting of a simulated cohort of 100 GBMs was chosen in this case to reflect the size of the patient cohort detailed in Section 4.2 and highlight challenges that may be encountered when inferring the model parameters for the patient dataset in the following section of this thesis.

The invasive advantages, ν_E^D and ν_P^D of the two amplified sub-populations are also inferred to some degree by the ABC-SMC inference algorithm. From Fig. 4.14(c), we see the region of highest density of accepted parameter values in the final population is in the bottom left hand corner of the plot, indicating that smaller values of ν_E^D and ν_P^D fit the synthetic data better. This effect is also observed in the scatter plot in Fig. 4.12(d), where it is clear that the majority of particles accepted into the final population in the ABC-SMC algorithm are those with smaller invasive advantages of the EGFR and PDGFRA amplified sub-populations. Thus, the inference results are consistent with the knowledge that these populations had no invasive advantage over the non-amplified population of tumour cells in the synthetic data.

Finally, we examine the inference results for the interaction parameters, α_{EP} and α_{PE}. The results, shown in Fig. 4.14(a), indicate that the most likely parameter values are those with α_{PE} around zero and small positive values of α_{EP}. The mean parameter values for α_{PE} and α_{EP} in the final population of accepted particles are 0.01 and 1.19, respectively. Furthermore, when plotting the results in α_{EP}–α_{PE} space, a relationship between these parameters becomes clear. From Fig.s 4.12(e) and 4.14(a), we see that competitive interactions, where both α_{EP} and α_{PE} are negative, are excluded from the final population of particles. Similarly, strongly cooperative interactions, where both parameters are larger than approximately 2.5, are also excluded. Instead, the cloud of accepted particles in the final population covers a range of other interaction types, including weakly cooperative, commensalism, parasitism of one population on another and neutralism, which we note was the true interaction type used to generate the synthetic data. Interestingly, this cloud is also almost symmetric along the line $\alpha_{EP} = \alpha_{PE}$

and its shape appears to suggest a negative correlation between the values of these parameters in the accepted particles. We saw from the sensitivity analysis in the previous chapter that opposite signs of α_{EP} and α_{PE} will benefit one amplified population, while having a negative effect on the other. We propose that this is the dominant effect of the interaction parameters in this case, as the values of these parameters in the final population of accepted particles vary approximately along the line $\alpha_{EP} = -\alpha_{PE} + C$, where $C \approxeq 1$. This line has a positive intercept on the α_{EP} and α_{PE} axes, as we know that these parameters positively affect the proportion of tumours that are amplified in both genes from our previous sensitivity analysis, which is the second largest mean proportion observed in the synthetic data in Fig. 4.8. Where these parameter values lie along this line will depend on the values of the other parameters in the accepted particles.

4.5 Parameter inference for the patient data

In this section, we now implement the ABC-SMC algorithm to infer the ten parameters of the model detailed in Table 4.1 for the patient cohort data shown in Fig. 4.4. As this data only tells us the amplification patterns observed in tumour biopsies and we look at the amplification levels across whole tumours in our simulated data, we make the assumption that the proportions of amplified and non-amplified biopsies for each tumour in the patient cohort is representative of the proportion of the whole tumour that is amplified and not amplified in each gene. In this way, we are able to infer the model parameters in the same way as for the synthetic dataset examples shown in the previous section.

Unlike for the synthetic data, the model parameters that best fit the patient data are unknown so the prior distributions for each parameter are now chosen based on some prior knowledge and assumptions about the populations of tumour cells. As EGFR and PDGFRA amplification are considered to be driver mutations in GBM tumour cells [107], playing major roles in driving the growth of GBMs, we assume that the parameters determining the proliferative and invasive advantages of the amplified sub-populations have a minimum value of 1. A value less than 1 would mean that EGFR or PDGFRA amplification reduces the proliferative or invasive ability of tumour cells, which contradicts the knowledge that they are driver mutations. We also assume that the maximum proliferative and invasive advantage that amplification in either these genes affords the tumour cells is twice the rate of proliferation and invasion of non-amplified tumour cells. Thus, the prior distributions for these parameters are taken to be uniform over

these ranges, namely

$$\nu_E^\rho \sim U(1,2), \tag{4.18}$$
$$\nu_P^\rho \sim U(1,2), \tag{4.19}$$
$$\nu_E^D \sim U(1,2), \tag{4.20}$$
$$\nu_P^D \sim U(1,2). \tag{4.21}$$

In Chapter 3, we assumed the interactions of the amplified sub-populations were symmetric and to the same degree, i.e. $\alpha_{EP} = \alpha_{PE}$, and explored the model behaviour with three different types of interaction; these were strong competition ($\alpha_{EP} = \alpha_{PE} = -5$), neutralism ($\alpha_{EP} = \alpha_{PE} = 0$) and strong cooperation ($\alpha_{EP} = \alpha_{PE} = 5$). In this work, we no longer assume that these parameters must be equal and take the prior distributions to be uniform distributions over these ranges. Thus, the prior distributions for α_{EP} and α_{PE} are defined as,

$$\alpha_{EP} \sim U(-5,5), \tag{4.22}$$
$$\alpha_{PE} \sim U(-5,5). \tag{4.23}$$

Finally, the prior distributions for each of the parameter pairs (k_E, θ_E) and (k_P, θ_P) are taken to be joint uniform distributions over the region S_i, where S_i is defined as

$$S_i = \{k_i, \theta_i \in \mathbb{R}^2 : 0.001 \leq k_i \leq 200,\ 0.001 \leq \theta_i \leq 5000,\ k_i\theta_i < 5000\}, \tag{4.24}$$

for $i = \{E, P\}$. As we have no prior knowledge about these parameters individually in relation to the patient cohort data, it is difficult to know what the prior distributions should be taken as. However, in Chapter 3 we explored introduction times for the amplified populations ranging from tumour sizes of 300 to 900cells/mm^2, so we now explore a much larger range of mean introduction times, up to tumour sizes of 5100cells/mm^2. Defining S_i in this way with large ranges of possible values of k_i and θ_i, means that a wide variety of gamma distribution shapes are possible for a large range of mean introduction tumour sizes.

We now implement the ABC-SMC algorithm as described in Section 4.4.1 with $Q+1 = 10$ populations of $R = 1000$ sampled particles to infer the unknown parameter vector $(\nu_E^\rho, \nu_P^\rho, \nu_E^D, \nu_P^D, k_E, \theta_E, k_P, \theta_P, \alpha_{PE}, \alpha_{EP})$ for the patient data in Fig. 4.4. As the patient cohort data is comprised of biopsies from 55 tumours, we produce a simulated cohort of 55 GBMs for each particle sampled throughout the inference, so the patient and simulated cohort sizes are the same. As previ-

ously, the tolerance vectors are generated adaptively as the inference algorithm progresses and, during this implementation, the values achieved are

$$\epsilon^M = (0.5548, 0.4170, 0.3371, 0.2719, 0.2154, 0.1723, 0.1401, 0.1155, 0.0964)$$

and

$$\epsilon^{SD} = (0.4638, 0.4113, 0.3766, 0.3509, 0.3263, 0.3043, 0.2852, 0.2672, 0.2513).$$

We note that the final tolerance level achieved for the distances between the mean data points from the patient cohort dataset and the cohorts simulated with the sampled particles, at $\epsilon_9^M = 0.0964$, is similar to the final tolerance level reached when the ABC-SMC algorithm was applied to the synthetic dataset in Section 4.4.3. This is promising as it indicates we are able to fit the model to these data points from the patient cohort data as well as to a synthetic dataset produced using the model. The final value of $\epsilon_9^{SD} = 0.2513$, however, is much larger, indicating that the final population of accepted particles do not produce simulated cohorts with a similar spread of proportion data to that observed in the patient cohort data. We note that the initial tolerance level in the vector ϵ^{SD} is much larger than the initial level observed when inferring the parameters for the synthetic dataset in the previous section. Thus, as the initial population of particles produces data further away, it may be the case that more iterations of the ABC-SMC algorithm are needed to achieve a similar final tolerance level for ϵ_Q^{SD}. As each successive value of ϵ_i^{SD} continues to steadily decrease for $i = 1, ..., 9$, this may be attainable if the number of populations sampled were increased, which could be explored in future work.

The results for this implementation of the ABC-SMC algorithm to infer the unknown model parameters for the patient cohort data shown in Fig. 4.4 are presented in Fig.s 4.16, 4.17 and 4.18.

Looking at plots (a) and (b) of Fig. 4.16 first, we observe that the ABC-SMC algorithm has been able to infer some information about the shape and scale parameters of the gamma distributions for N_E and N_P. We observe that the values of k_E in the final population of accepted particles with the greatest density lie mostly between 2 and 4, with a mean value of 2.56. Meanwhile, the values for the θ_E parameter produce a much wider distribution, with a wide variety of values accepted ranging from 114.86 to 2482.81 and a mean value of 873.89. The kernel density estimate plot for the k_P and θ_P parameters shown in (b) is a different shape to the plot in (a); namely, it is wider in the k_P direction

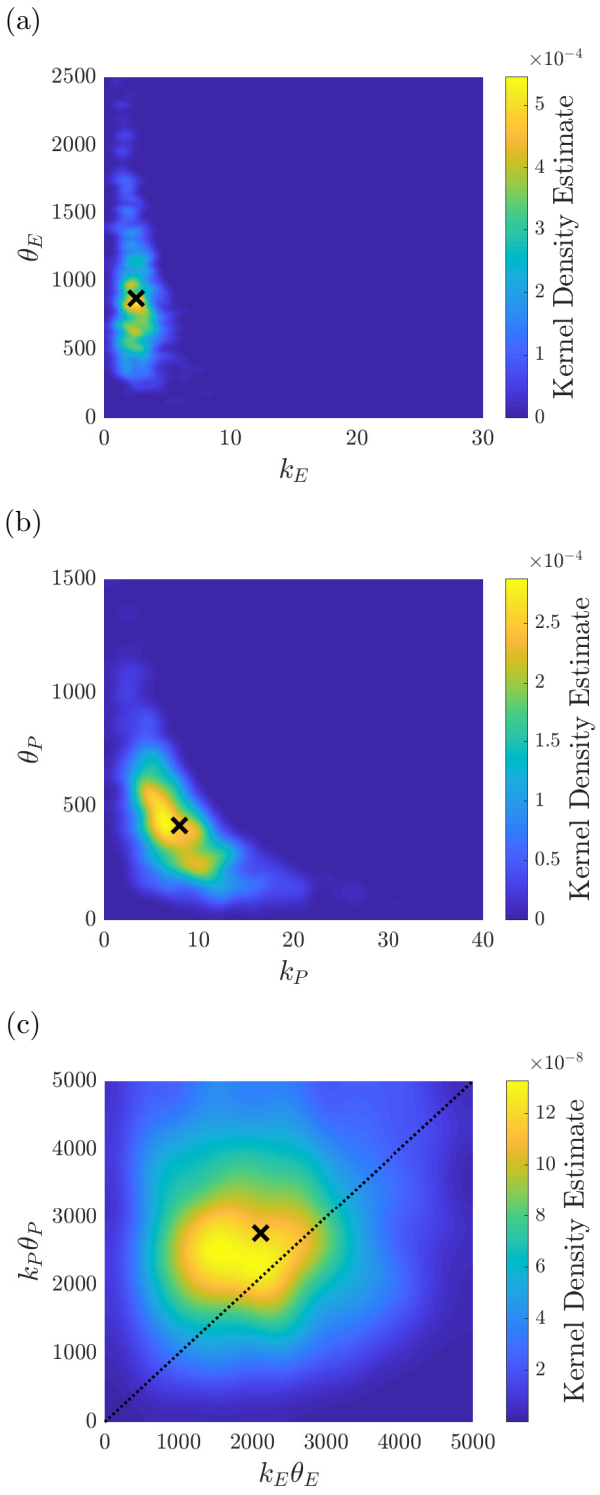


Figure 4.16: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters for the patient data shown in Fig.4.4. Each plot shows kernel density estimates of the final population of accepted parameter particles in (a) $k_E - \theta_E$, (b) $k_P - \theta_P$ and (c) $k_E\theta_E - k_P\theta_P$ parameter space. The mean of the final population of accepted particles is shown by the black '$\times$' on each plot. On plot (c), the line $k_E\theta_E = k_P\theta_P$ is shown by the black dotted line.

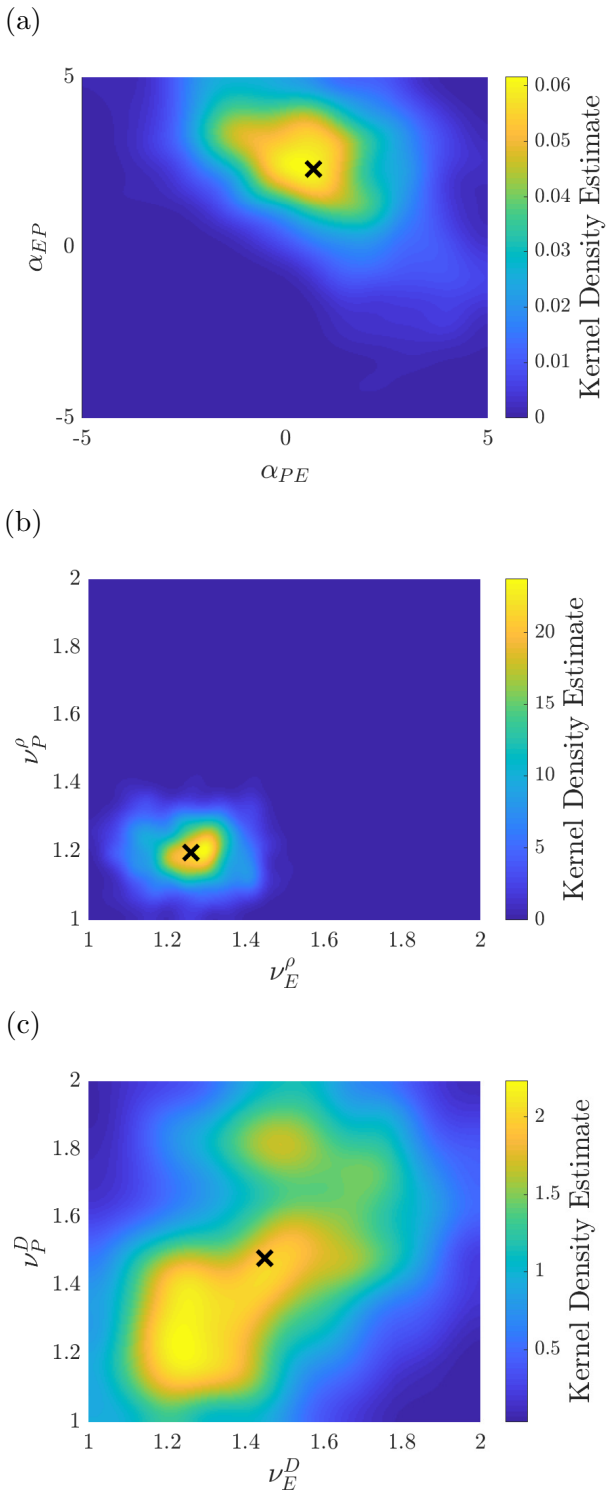


Figure 4.17: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters for the patient data shown in Fig.4.4. Each plot shows kernel density estimates of the final population of accepted parameter particles in (a) $\alpha_{PE} - \alpha_{EP}$, (b) $\nu_E^\rho - \nu_P^\rho$ and (c) $\nu_E^D - \nu_P^D$ parameter space. The mean of the final population of accepted particles is shown by the black '$\times$' on each plot.

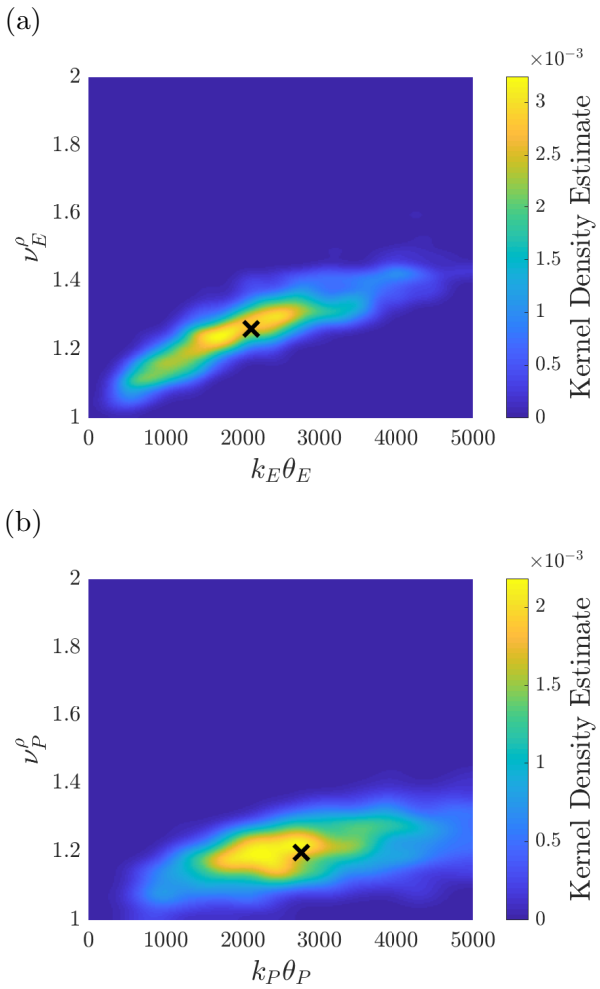


Figure 4.18: Results from an implementation of the ABC-SMC algorithm to infer the unknown model parameters for the patient data shown in Fig.4.4. Each plot shows kernel density estimates of the final population of accepted parameter particles in (a) $k_E\theta_E - \nu_E^\rho$ and (b) $k_P\theta_P - \nu_P^\rho$ parameter space. The mean of the final population of accepted particles is shown by the black '$\times$' on each plot.

and shorter in the θ_P direction, with the greatest density of values lying below and to the right of that in plot (a). The mean values of k_P and θ_P in the final population of accepted particles are 7.98 and 414.98, respectively.

Following on from this, we see that a wide range of mean values of the gamma distributions for N_E and N_P are in the final population of particles. Figure 4.16(c) reveals that the majority of the high density region of particles lies to the left of the line $k_E\theta_E = k_P\theta_P$, i.e. where the EGFR amplified population of cells are introduced at earlier mean times into the growing tumour than the PDGFRA amplified population. The mean values of $k_E\theta_E$ and $k_P\theta_P$ are 2117.5 and 2764.9, respectively.

The next inference results to examine are for the interaction parameters, α_{EP} and α_{PE}. In Fig. 4.17(a), we observe that highest density region for the values

of these parameters in the final population of accepted particles lie in the top half of the graph close to the centre, where the presence of EGFR amplified cells have a positive effect on the growth of PDGFRA amplified cells and the effect of PDGFRA amplified cells on the EGFR amplified sub-population is less strong. The mean values for these parameters, α_{EP} and α_{PE} in the final population of accepted particles are 2.31 and 0.70, respectively. The accepted particles in the final population are distributed over the four quadrants of the scatter plot as follows: 346 particles have $\alpha_{EP} > 0$ and $\alpha_{PE} \leq 0$; no particles have $\alpha_{EP} \leq 0$ and $\alpha_{PE} \leq 0$; 92 particles have $\alpha_{EP} \leq 0$ and $\alpha_{PE} > 0$; and 562 particles have $\alpha_{EP} > 0$ and $\alpha_{PE} > 0$. The numbers of accepted particles in each quadrant of the scatter plot suggest that a cooperative relationship between the two populations is most likely, where the amplified cell populations benefit from the presence of one another. This is followed by parasitism between the two populations as the next most likely interaction type, where PDGFRA amplified cells benefit from the presence of the EGFR amplified cells, while negatively affecting their growth. It is clear, however, that the inference results for do not suggest a competitive relationship between the EGFR and PDGFRA amplified sub-populations, as no particles were accepted into the final population of particles with α_{EP} and α_{PE} both negative.

The plot in Fig. 4.17(b) shows the density of ν_E^ρ and ν_P^ρ values in the final population of accepted particles. From this we observe that all values above 1.6 have been excluded, indicating that the amplified sub-populations do not have a very large proliferative advantage over the non-amplified tumour cells. We observe that the region of highest density for the proliferative advantages of the amplified cells suggests that EGFR amplification affords cells a slightly higher proliferative ability than PDGFRA amplified cells, with ν_E^ρ and ν_P^ρ having mean values of 1.26 and 1.20, respectively, in the final population of accepted particles.

The estimated densities of the invasive advantage parameters, ν_E^D and ν_P^D accepted into the final population are shown in Fig. 4.17(c). From this we observe that, although particles are accepted with parameter values spanning the width of the prior distribution, a clear peak in density has formed between values of 1.2 and 1.4 for both parameters. This suggests that EGFR and PDGFRA amplification may afford tumour cells a greater ability to invade tissue than non-amplified cells. The density distribution in both the ν_E^D and ν_P^D directions look very similar, with mean values at 1.45 and 1.48, indicating that cells amplified in EGFR and PDGFRA are similarly invasive.

As noted in the previous section, the proliferative advantage of amplified cells

and their mean introduction time into the growing tumour are correlated, as the density plots in Fig.4.18(a) and (b) also demonstrate in this case. From scatter plot (a), we see that particles with earlier mean introduction times are accepted when the proliferative advantage of EGFR amplified cells is small, while those with later introduction times are accepted with larger proliferative advantages; a similar pattern is also observed for the PDGFRA amplified cells in Fig. 4.18(b). Finally, we observe that the estimated density of the final population of accepted particle values in $k_E\theta_E - \nu_E^\rho$ space lies above and to the left of that in $k_P\theta_P - \nu_P^\rho$ space, illustrating, again, that the EGFR amplified cells have a proliferative advantage over and are introduced earlier into the growing tumour than the PDGFRA amplified cells.

4.6 Discussion

In this chapter, we have used our mathematical model describing the co-evolution of three distinct tumour cell sub-populations to infer the dynamics and nature of interactions between EGFR and PDGFRA amplified populations in glioblastomas from a set of patient data.

The patient dataset highlights the importance of sampling from multiple regions of a tumour as large variations in amplification levels of the EGFR and PDGFRA genes can be observed between different regions of individual tumours, which would not be identified through more common sampling techniques where only one sample is analysed per tumour. In particular, a patient example was illustrated (see Fig. 4.2), where 8 biopsy samples were analysed and heterogeneity in EGFR and PDGFRA amplification was observed across the tumour region. In addition, the number of tumours found to have at least one biopsy sample amplified in EGFR, PDGFRA or both genes is much higher in this dataset than in other reported data [107, 125], further highlighting the importance of multi-region sampling techniques as conventional sampling techniques may under-represent the true number of tumours having amplification of these genes, which may have implications when identifying patients that will benefit from targeted therapies.

As a wide variety of amplification patterns are observed in the patient dataset (see Fig. 4.4a), we assumed that certain model parameters had distributions associated with them in order to reflect the role that natural variation and the knowledge that mutations occur randomly may be playing in this. In this way, a variety of amplified proportions are observed in a simulated cohort of GBMs, echoing that observed in the patient dataset. In particular, we assumed that

the tumour size at which the amplified sub-populations were introduced into the growing tumour, N_E and N_P, were gamma distributed, with associated shape, k_E and k_P, and scale parameters, θ_E and θ_P, respectively. We explored the effect that varying these shape and scale parameters has on the proportions and mean proportions of amplified tumours in a simulated cohort of GBMs and found that smaller values of the shape parameters, k_E and k_P, produced a wider spread of proportions and had some impact on the mean proportions observed (see Fig.s 4.6 and 4.7).

Next, we tested whether these new model parameters—k_E, k_P, θ_E and θ_P—could be inferred correctly using the ABC-SMC algorithm when all other model parameters are known for a synthetic dataset. We tested this on synthetic datasets with two different sets of values for the unknown parameters, the first with $k_E = k_P = 10$ and $\theta_E = \theta_P = 20$ and the second with $k_E = k_P = 1$ and $\theta_E = \theta_P = 1000$ (see Fig.s 4.9 and 4.11). We found that the parameters for the first synthetic dataset were inferred better than for the second, where the scale parameters $\theta_E = \theta_P = 1000$ were inferred quite poorly. We found in both cases that we were able to infer the means of the gamma distributions well, with the mean for the first dataset also being inferred slightly better than the second. It is likely that the inference algorithm performed better at inferring the parameters for the first set of synthetic data than the second due to the increased variation between sets of synthetic data generated with the second set of parameters. This variation presents a challenge for inference as a single simulated cohort may not provide much information about the mean dynamics of the system and it results in a lower particle acceptance rate. Increasing the number of simulated tumours in the synthetic dataset and for each particle sampled would help to improve this, although the simulated cohorts were chosen to contain 100 GBMs in order to test the algorithm on a dataset of similar size to the patient dataset, to which the inference algorithm is later applied. However, this is something that could be explored in future work as it may highlight the importance of a larger patient dataset if more data were to become available in future. Other factors that may help to improve the inference results are increasing the number of particles in each population and also the number of populations. Although, we note in this case that we were able to achieve final tolerance levels in line with the variation observed between simulated cohorts generated with the same parameters so increasing the number of populations is unlikely to yield much improvement without the size of the data and simulated cohorts being increased. Increasing these three factors would help to improve the inference results to some degree, however it would also

greatly increase the computational burden of implementing the algorithm so it would be important to consider this in future work. Each of the implementations of the ABCSMC algorithm in this work took between approximately 5-10 days to run, depending on the numbers of particles and populations sampled and the number of parameters inferred.

The inference algorithm was then used to test whether all 10 model parameters could be inferred for a synthetic dataset. Because of the increased dimensionality of this problem, the number of particles in each population was increased to reflect this, however, as noted before, this could be increased further to improve our results. Nevertheless, applying the ABC-SMC algorithm to this problem did allow us to infer some information about the model parameters, some more successfully than others. In particular, we note that in this case the means of the gamma distributions for the size of tumour when the amplified sub-populations are introduced were not inferred well in this case, since the final population of accepted particles did not appear to be converging around the true parameter values as seen previously. This is likely to be as a result of the increased dimensionality of the inference and in Fig.s 4.12(g) and (h) it was shown that there is a relationship between accepted parameter values for the mean introduction times and proliferative advantages of the EGFR and PDGFRA amplified sub-populations. Specifically, sampled particles with later introduction times for the amplified sub-populations are only accepted for larger values of proliferation advantages. Intuitively, this makes sense as a more proliferative population of tumour cells will be able to reach the same size as a less proliferative population that started growing earlier. Indeed, this is consistent with the sensitivity analysis presented in Chapter 3, where it was observed that the proportions of simulations with only EGFR and only PDGFRA amplified were both strongly correlated to proliferation advantages and the timing of mutations, with the effects being reflected for each output.

Finally, we inferred the model parameters for the patient cohort data detailed in Section 4.2, where we were able to gain some insight into the dynamics of EGFR and PDGFRA amplified sub-populations in glioblastomas. In particular, the results suggest that EGFR amplification is a mutation event typically occurring at mean earlier times than PDGFRA amplification in the development of GBMs, consistent with the findings of a study reconstructing the phylogeny of these tumours [108]. The results also indicate that the amplified sub-populations may have a slight proliferative advantage over the non-amplified tumour cells, as the mean parameter values for these advantages were greater than one in both

cases, with the EGFR amplified population having the greater mean proliferation advantage. We note that particles with large proliferative advantages were not accepted in the inference, indicating that amplification in either of these genes does not afford cells a large increase in proliferative ability; we note that this suggests the proliferation advantages of 50% explored in Chapter 3 may have been too large, as speculated in the discussion in Section 3.5. These results also illustrate the correlation between proliferation advantages and the timing of mutations, with later mutation events being correlated to larger proliferative advantages for both of the amplified sub-populations, consistent with the results from the sensitivity analysis presented in Chapter 3.

Data at multiple time points, such as biopsies sampled from primary and recurrent tumours, would be needed to provide us with the dynamic information in order to further determine estimates and improve the identifiability of such related parameters. This would present challenges, however, as treatment effects would also need to be considered. Alternatively, radiogenomic maps showing predicted regions of EGFR and PDGFRA amplification, such as those in [42] discussed in Section 1.1.1, could provide data from multiple time points in a non-invasive manner by utilising multiple pre-treatment MRIs to create maps a different time points. Thus, instead of having only biopsy data from a single time point, these maps would provide us with information about the amplified proportions of the entire tumour region at multiple time points. This would enable us to infer better estimates of the model parameters and improve the identifiability of those model parameters found to be related to one another, such as the proliferative advantages and timing of mutations. Utilising radiogenomic maps in this way would also remove the need to consider treatment or surgical effects on the observed amplification patterns, as would be necessary to consider with biopsy data sampled at two time points. Another approach would be to use *in vitro* single-cell cultures, which may be able to shed some light on the relative proliferative abilities of amplified and non-amplified GBM tumour cells and provide an insight into their *in vivo* dynamics.

The inference results also suggest that amplification in the EGFR and PDGFRA genes each result in more invasive GBM tumour cells to a similar degree. EGFR amplification has been shown to promote invasion in GBMs [124], while EGFR amplified tumours have also been shown to be more migratory than non-amplified tumours [81] and results from a recent study support an association between EGFR amplification in cells and faster migratory behaviour of cells in GBM slice cultures [80]. Thus, our finding that EGFR amplified cells have an invasive ad-

vantage over non-amplified cells is consistent with the literature. In their work, Parker et al. [80] observe a wide variety of migration speeds in EGFR amplified GBM slice cultures and postulate that this may be explained by the varying levels of EGFR amplification they observe. We also observe varying degrees of EGFR amplification in our biopsy data (see Fig. 4.3), however our model does not take this into consideration; cells are either amplified in the EGFR gene or they are not. Introducing variable invasive abilities and varying degrees of EGFR amplification in our model formulation may help to improve our results further and provide a better fit to the data than achieved here; this is something that could be explored in future work. Meanwhile, amplification of PDGFRA in GBMs is less well studied and we have been unable to find evidence in the literature to support or contradict our findings that PDGFRA amplification may increase the invasive ability of glioblastoma cells, similarly to EGFR amplification. Thus, it would be interesting to see if similar behaviour is observed in PDGFRA amplified GBM cells in *in vivo* studies, such as those in [80].

Finally, our inference results provide some insight into the nature of interactions between EGFR and PDGFRA amplified sub-populations. In particular, our model suggests that cooperation is the most likely interaction type between the amplified sub-populations, with each benefiting from the presence of the other, resulting in their increased proliferation and contributing to overall increased tumour growth. If the EGFR and PDGFRA amplified cells are indeed interacting in this way, targeting either sub-population will also have a negative impact on the other population by removing their promoter, thus potentially increasing the therapeutic benefits in glioblastomas where populations of both EGFR and PDGFRA amplified cells are present. Our finding that cooperativity is the most likely interaction type between these amplified cells is consistent with studies in the literature that point to cooperation between these sub-populations in glioblastomas [16, 107, 123], although we do note that more research is needed in this area as the mechanism by which PDGFRA amplified and EGFR amplified cells may interact in a cooperative manner is unclear and has not been well studied.

The second most likely interaction type in our results is parasitism of the PDGFRA amplified cells on the EGFR amplified sub-population. If the cells are interacting in this way, then EGFR amplified cells promote PDGFRA amplified cells, while being negatively affected by them. Such interactions would have implications for targeted therapies, as only targeting PDGFRA amplified cells would be beneficial for the EGFR amplified cells, allowing them to proliferate more freely once the PDGFRA amplified population is removed. Alternatively,

targeting EGFR amplified cells first would remove the EGFR cells promoting the PDGFRA amplified population, which could then be targeted with a different therapy. We note that if there is uncertainty between a cooperative interaction type or parasitism of the PDGFRA cells on the EGFR amplified sub-population, such a treatment regimen would work well in both cases, since the PDGFRA amplified cells benefit from the presence of the EGFR amplified sub-population in both instances.

Thus, our results suggest that targeting the EGFR amplified cells with therapy first, followed by targeting the PDGFRA amplified sub-population would be the best course of action to take. Over 90% of the accepted particles in our inference results say that EGFR amplified cells promote the growth of PDGFRA amplified cells, whereas the influence of the PDGFRA amplified cells on the EGFR amplified cells is less clear, although cooperation is the most likely. Our results also suggest that competition between these amplified sub-populations is unlikely as no particles with competitive interaction parameters were accepted into the final population.

While this work has provided some insights into the interactions between and dynamics of EGFR and PDGFRA amplified sub-populations in glioblastomas, more work needs to be carried out. As this is a complex biological problem, with several factors influencing the amplification patterns observed in our model, more data will be needed in order to improve our results further to make them more robust, improve our confidence in the results and improving the ability to distinguish parameter estimates more clearly. This could be through increasing the size of the patient cohort data further or using a combination of single- and mixed-cell cultures and mathematical modelling, which may be able to identify the nature of interactions *in vitro* and provide some useful insights into the co-evolution of EGFR and PDGFRA sub-populations *in vivo*. Ideally, biopsy information at multiple time points would be available, allowing us to compare our model simulations to the patient data at more than one time point. However, this is unlikely to be possible without at least needing to account for treatment effects, adding further complexity to the problem. Alternatively, the emerging field of radiogenomics may have a role to play here, as predicted distributions of EGFR and PDGFRA amplification could be obtained at multiple time points from imaging data non-invasively.

We also note that in this work we have avoided modelling single cell events in a macroscopic setting, by assuming that each of the EGFR and PDGFRA amplified sub-populations only become established within a tumour at most once

and introduce a small distribution of cells accordingly. It is possible, however, that multiple mutation events resulting in EGFR and PDGFRA amplification in cells could be occurring in an evolving tumour and it may be more appropriate to model such single cell events in a micro- or meso-scopic setting. However, since we are interested in population level patterns of amplification and are working with large numbers of cells, this would be computationally expensive. In order to better capture single cell events while avoiding a large computational burden, it may be appropriate to consider a hybrid-modelling approach, similar to that described by Smith and Yates [104]. Though in this work, we decided to represent a successful mutation event by introducing a small population of cells in our continuum PDE model and leave such considerations for future work.

Chapter 5

Conclusion

In this thesis, we have used a novel mathematical model describing the growth of three distinct sub-populations to explore the nature of interactions between EGFR and PDGFRA amplified sub-populations in glioblastomas. This work was motivated by the failure of therapies to treat GBMs, in particular the failure of targeted therapies, where it is thought that the heterogeneous nature of these tumours may be an underlying cause. Genetically distinct sub-populations may coexist within a single tumour and interact in such a way that enables them to evade therapy or play an important role in tumour progression. In this work, we have focussed on studying the dynamics of EGFR and PDGFRA amplified sub-populations. These are two commonly occurring cell-types in GBMs and there are some studies suggesting that they may be interacting in a cooperative manner, although this requires further study. A particular challenge to overcome when studying the nature of dynamics between these sub-populations in GBMs is the lack of dynamic information available. Image-localised biopsies provide important genetic and spatial information about the distribution and co-occurrence of EGFR and PDGFRA amplified sub-populations, however, this information is static and it is difficult to extract dynamic information from. Therefore, in this work, we have utilised our mathematical model to provide insight into the complex dynamics of these amplified tumour cells.

In Chapter 2, we presented our novel mathematical model of the growth of three distinct tumour cell sub-populations in GBMs; a sub-population amplified in the EGFR gene, another amplified in the PDGFRA gene and a third population amplified in neither gene. We explored the model behaviour under a variety of interaction types and found that the nature of interactions resulted in some interesting model dynamics. In particular, we found that a proliferative advantage of one amplified sub-population over the other may not always actually be an

advantage for those cells, as there are scenarios where the less aggressive tumour cell population will become the dominant one (see Fig. 2.3). We then explored travelling wave solutions to the model through numerical simulations and found some conditions for their existence. We found that a variety of travelling waves can emerge, with the type of wave we see depending on the model parameters and initial conditions. These observations could motivate the design of *in vitro* experiments, whereby the initial configuration of EGFR and PDGFRA amplified cells is varied and the patterns of invasion that emerge are studied. This could provide information about the type of interactions occurring between these glioblastoma cell sub-populations and provide validation of our model.

We also showed that EGFR and PDGFRA amplified cells can co-exist under a variety of different interaction types. In particular, we showed that a competition co-existence state exists when there is strong competition between these sub-populations and demonstrated that regions of co-existence can occur under a variety of interaction assumptions (see Fig. 3.2). This indicates that the observation of EGFR and PDGFRA cells co-existing in the same tumour region does not necessarily imply that these cells are cooperating in some way, as we may intuitively expect, and further evidence is needed.

Following on from this, in Chapter 3, we conducted an *in silico* investigation, where we compared the amplification levels of EGFR and PDGFRA in our simulations to those in a preliminary dataset of image-localised biopsies from primary GBMs. We found that factors relating to selection advantages and the phylogeny of these tumours, in addition to the interaction type between the EGFR and PDGFRA amplified cells, influenced the balance of populations we see in a simulated tumour. These additional factors make determining the nature of interactions between the amplified sub-populations in GBMs more challenging, as they add complexity to the problem. Nevertheless, we were able to gain important insights into the dynamics of EGFR and PDGFRA amplified sub-populations through the investigation. Importantly, we found that our results suggested that EGFR amplification is a mutation occurring earlier than PDGFRA amplification in the growth of GBMs (see Fig. 3.5), consistent with the findings from a study of the phylogeny of GBMs [108]. Our findings in this chapter also revealed that strong cooperative interactions between the amplified sub-populations did not produce amplification patterns that compared as well to the levels observed in the preliminary patient data as competitive and neutral interactions did.

Finally, in Chapter 4, we presented the complete patient dataset of image-localised biopsies and used inference techniques to gain insight into the dynamics

of EGFR and PDGFRA amplified sub-populations in GBMs. We highlighted the intra-tumoural heterogeneity of amplification in these genes through presenting a patient example (see Fig. 4.2) and the inter-tumoural heterogeneity as a variety of amplification patterns are observed across the patient cohort (see Fig. 4.4a). To reflect this inter-tumoural heterogeneity, we assumed that certain model parameters had associated probability distributions, so that a variety of amplification patterns are observed when simulating a cohort of GBMs, echoing the patient data. To test whether we could recover the model parameters, we generated some synthetic datasets and employed an ABC-SMC inference algorithm, where we found that some model parameters were inferred better than others and that the level of variation in the model presented a challenge. Nevertheless, we then inferred the model parameters for the patient data, where we were still able to gain some insight into the dynamics of EGFR and PDGFRA amplified sub-populations. Some key suggestions from the results were that: amplification in each of these genes confers cells slight selection advantages over non-amplified cells; EGFR amplification is a mutation that is likely to occur, in general, earlier than PDGFRA amplification in the growth of GBMs, consistent with our previous findings and another study [108]; and cooperation is the most likely interaction type between these amplified sub-populations. Taking these findings into account, we suggested that a treatment regimen whereby EGFR amplified cells are targeted with therapy first, followed by the targeting of PDGFRA amplified cells would be likely to result in the best results for patients with GBM tumours where both amplified sub-populations are present.

Thus, despite only having limited data from a single time point, we have been able to use our mathematical model to provide an insight into the dynamics of EGFR and PDGFRA amplified sub-populations in GBMs and the nature of interactions between them. As discussed in Section 4.6, this work has some limitations that could be improved on and there are several directions for future work, some of which we will briefly mention here. A simple next step, would be to re-run the inference as more patient data becomes available to improve our confidence in the results. It may also be of interest to incorporate different levels of amplification in the EGFR and PDGFRA genes into our model; currently cells are either amplified or not amplified in each gene, whereas a recent study suggested that the varying levels of migrative ability of cells may be a result of varying levels of EGFR amplification [80], so this could be something to explore in future work that may improve our results. Finally, it would be particularly interesting to utilise our model with predictions of EGFR and PDGFRA amplification from

radiogenomic maps. Firstly, it would be interesting to compare inference results using the predicted amplification patterns from the maps to those in the patient biopsy data. Further to this, as these maps provide information on predicted distributions of amplified cell populations non-invasively from imaging data, this could be used to fit our model to multiple time points, which could help to further determine model parameters, particularly where correlations exist between them (see Fig.s 4.18a and 4.18b) and their identifiability when using data from only a single time point is a challenge.

Determining the nature of interactions between EGFR and PDGFRA amplified sub-populations in GBMs is a complex biological problem, with potential clinical implications for targeted therapies and aiding a deeper understanding of how these aggressive tumours evolve and evade treatments. To be able to further untangle the influence from and the nature of interactions from the effects of other factors influencing the patterns of EGFR and PDGFRA amplification observed in GBMs, more work needs to be done. This will likely involve a combination of *in silico*, *in vitro* and, possibly, *in vivo* studies to characterise the dynamics of these sub-populations, as well as cell populations harbouring other mutations commonly occurring in GBMs. In this work, however, we have demonstrated that a simple mathematical model can prove a useful tool and provide some important insights and aid more understanding of these complex brain tumours.

Bibliography

[1] Alentorn, A., Marie, Y., Carpentier, C., Boisselier, B., Giry, M., Labussiere, M., Mokhtari, K., Hoang-Xuan, K., Sanson, M., Delattre, J.-Y., et al. Prevalence, clinico-pathological value, and co-occurrence of PDGFRA abnormalities in diffuse gliomas. *Neuro-oncology 14*, 11 (2012), 1393–1403.

[2] Alexander, A. L., Lee, J. E., Lazar, M., and Field, A. S. Diffusion tensor imaging of the brain. *Neurotherapeutics 4*, 3 (2007), 316–329.

[3] Alfonso, J., Talkenberger, K., Seifert, M., Klink, B., Hawkins-Daarud, A., Swanson, K., Hatzikirou, H., and Deutsch, A. The biology and mathematical modelling of glioma invasion: a review. *Journal of The Royal Society Interface 14*, 136 (2017), 20170490.

[4] Altrock, P. M., Liu, L. L., and Michor, F. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer 15*, 12 (2015), 730.

[5] An, Z., Aksoy, O., Zheng, T., Fan, Q.-W., and Weiss, W. A. Epidermal growth factor receptor and EGFRvIII in glioblastoma: signaling pathways and targeted therapies. *Oncogene 37*, 12 (2018), 1561–1575.

[6] Baldock, A. L., Ahn, S., Rockne, R., Johnston, S., Neal, M., Corwin, D., Clark-Swanson, K., Sterin, G., Trister, A. D., Malone, H., et al. Patient-specific metrics of invasiveness reveal significant prognostic benefit of resection in a predictable subset of gliomas. *PloS One 9*, 10 (2014), e99057.

[7] Baldock, A. L., Yagle, K., Born, D. E., Ahn, S., Trister, A. D., Neal, M., Johnston, S. K., Bridge, C. A., Basanta, D., Scott, J., et al. Invasion and proliferation kinetics in enhancing gliomas predict IDH1 mutation status. *Neuro-Oncology 16*, 6 (2014), 779–786.

[8] Barajas Jr, R. F., Hodgson, J. G., Chang, J. S., Vandenberg, S. R., Yeh, R.-F., Parsa, A. T., McDermott, M. W., Berger, M. S., Dillon, W. P., and Cha, S. Glioblastoma multiforme regional genetic and cellular expression patterns: influence on anatomic and physiologic mr imaging. *Radiology 254*, 2 (2010), 564–576.

[9] Bergers, G., and Benjamin, L. E. Angiogenesis: tumorigenesis and the angiogenic switch. *Nature Reviews Cancer 3*, 6 (2003), 401.

[10] Blower, S. M., and Dowlatabadi, H. Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example. *International Statistical Review/Revue Internationale de Statistique* (1994), 229–243.

[11] Blume-Jensen, P., and Hunter, T. Oncogenic kinase signalling. *Nature 411*, 6835 (2001), 355–365.

[12] Bonavia, R., Cavenee, W. K., Furnari, F. B., et al. Heterogeneity maintenance in glioblastoma: a social network. *Cancer Research 71*, 12 (2011), 4055–4060.

[13] Borad, M. J., Champion, M. D., Egan, J. B., Liang, W. S., Fonseca, R., Bryce, A. H., McCullough, A. E., Barrett, M. T., Hunt, K., Patel, M. D., et al. Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma. *PLoS Genetics 10*, 2 (2014), e1004135.

[14] Burger, M., Di Francesco, M., Pietschmann, J.-F., and Schlake, B. Nonlinear cross-diffusion with size exclusion. *SIAM Journal on Mathematical Analysis 42*, 6 (2010), 2842–2871.

[15] Burger, M., Friele, P., and Pietschmann, J.-F. On a reaction-cross-diffusion system modelling the growth of glioblastoma. *arXiv preprint arXiv:1710.03970* (2017).

[16] Chen, F., and Ding, L. Co-survival of the fittest few: mosaic amplification of receptor tyrosine kinases in glioblastoma. *Genome Biology 13*, 1 (2012), 1–3.

[17] Claes, A., Idema, A. J., and Wesseling, P. Diffuse glioma growth: a guerilla war. *Acta Neuropathologica 114*, 5 (2007), 443–458.

[18] Clatz, O., Sermesant, M., Bondiau, P.-Y., Delingette, H., Warfield, S. K., Malandain, G., and Ayache, N. Realistic simulation of the 3-D growth of brain tumors in MR images coupling diffusion with biomechanical deformation. *IEEE Transactions on Medical Imaging 24*, 10 (2005), 1334–1346.

[19] Corwin, D., Holdsworth, C., Rockne, R. C., Trister, A. D., Mrugala, M. M., Rockhill, J. K., Stewart, R. D., Phillips, M., and Swanson, K. R. Toward patient-specific, biologically optimized radiation therapy plans for the treatment of glioblastoma. *PloS One 8*, 11 (2013), e79115.

[20] Cunningham, J. J., Gatenby, R. A., and Brown, J. S. Evolutionary dynamics in cancer therapy. *Molecular Pharmaceutics 8*, 6 (2011), 2094–2100.

[21] De Leon, G., Singleton, K. W., and Swanson, K. R. Days gained: a simulation-based, response metric in the assessment of glioblastoma. *bioRxiv* (2018), 209056.

[22] Del Moral, P., Doucet, A., and Jasra, A. An adaptive sequential monte carlo method for approximate Bayesian computation. *Statistics and Computing 22*, 5 (2012), 1009–1020.

[23] Demerath, T., Simon-Gabriel, C. P., Kellner, E., Schwarzwald, R., Lange, T., Heiland, D. H., Reinacher, P., Staszewski, O., Mast, H., Kiselev, V. G., et al. Mesoscopic imaging of glioblastomas: are diffusion, perfusion and spectroscopic measures influenced by the radiogenetic phenotype? *The Neuroradiology Journal 30*, 1 (2017), 36–47.

[24] Demuth, T., and Berens, M. E. Molecular mechanisms of glioma cell migration and invasion. *Journal of Neuro-Oncology 70*, 2 (2004), 217–228.

[25] Drovandi, C. C., and Pettitt, A. N. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics 67*, 1 (2011), 225–233.

[26] Eisenhauer, E., Therasse, P., Bogaerts, J., Schwartz, L., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. New response evaluation criteria in solid tumours:

revised recist guideline (version 1.1). *European Journal of Cancer 45*, 2 (2009), 228–247.

[27] Filippi, S., Barnes, C. P., Cornebise, J., and Stumpf, M. P. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical Applications in Genetics and Molecular Biology 12*, 1 (2013), 87–107.

[28] Fisher, R. A. The wave of advance of advantageous genes. *Annals of Human Genetics 7*, 4 (1937), 355–369.

[29] Frieboes, H. B., Jin, F., Chuang, Y.-L., Wise, S. M., Lowengrub, J. S., and Cristini, V. Three-dimensional multispecies nonlinear tumor growth—II: tumor invasion and angiogenesis. *Journal of Theoretical Biology 264*, 4 (2010), 1254–1278.

[30] Furnari, F. B., Cloughesy, T. F., Cavenee, W. K., and Mischel, P. S. Heterogeneity of epidermal growth factor receptor signalling networks in glioblastoma. *Nature Reviews Cancer 15*, 5 (2015), 302–310.

[31] Gardner, W. J., Karnosh, L., McCLURE JR, C. C., and GARDNER, A. K. Residual function following hemispherectomy for tumour and for infantile hemiplegia. *Brain 78*, 4 (1955), 487–502.

[32] Gerlee, P., and Nelander, S. The impact of phenotypic switching on glioblastoma growth and invasion. *PLoS Computational Biology 8*, 6 (2012), e1002556.

[33] Gerlee, P., and Nelander, S. Travelling wave analysis of a mathematical model of glioblastoma growth. *Mathematical Biosciences 276* (2016), 75–81.

[34] Gilbert, M. R., Kuhn, J., Lamborn, K. R., Lieberman, F., Wen, P. Y., Mehta, M., Cloughesy, T., Lassman, A. B., DeAngelis, L. M., Chang, S., et al. Cilengitide in patients with recurrent glioblastoma: the results of NABTC 03-02, a phase II trial with measures of treatment delivery. *Journal of Neuro-Oncology 106*, 1 (2012), 147–153.

[35] Gu, S., Chakraborty, G., Champley, K., Alessio, A. M., Claridge, J., Rockne, R., Muzi, M., Krohn, K. A., Spence, A. M., Alvord Jr, E. C., et al. Applying a patient-specific bio-mathematical

model of glioma growth to develop virtual [18F]-FMISO-PET images. *Mathematical Medicine and Biology: A Journal of the IMA 29*, 1 (2012), 31–48.

[36] Hawkins-Daarud, A., Johnston, S. K., and Swanson, K. R. Quantifying uncertainty and robustness in a biomathematical model–based patient-specific response metric for glioblastoma. *JCO Clinical Cancer Informatics 3* (2019), 1–8.

[37] Holland, E. C. Glioblastoma multiforme: the terminator. *Proceedings of the National Academy of Sciences 97*, 12 (2000), 6242–6244.

[38] Hormuth, D. A., Al Feghali, K. A., Elliott, A. M., Yankeelov, T. E., and Chung, C. Image-based personalization of computational models for predicting response of high-grade glioma to chemoradiation. *Scientific Reports 11*, 1 (2021), 1–14.

[39] Hottinger, A. F., Stupp, R., and Homicsko, K. Standards of care and novel approaches in the management of glioblastoma multiforme. *Chinese Journal of Cancer 33*, 1 (2014), 32.

[40] Hu, L. S., Ning, S., Eschbacher, J. M., Baxter, L. C., Gaw, N., Ranjbar, S., Plasencia, J., Dueck, A. C., Peng, S., Smith, K. A., et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro-Oncology 19*, 1 (2016), 128–137.

[41] Hu, L. S., Ning, S., Eschbacher, J. M., Baxter, L. C., Gaw, N., Ranjbar, S., Plasencia, J., Dueck, A. C., Peng, S., Smith, K. A., et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro-Oncology 19*, 1 (2017), 128–137.

[42] Hu, L. S., Wang, L., Hawkins-Daarud, A., Eschbacher, J. M., Singleton, K. W., Jackson, P. R., Clark-Swanson, K., Sereduk, C. P., Peng, S., Wang, P., et al. Uncertainty quantification in the radiogenomics modeling of EGFR amplification in glioblastoma. *Scientific Reports 11*, 1 (2021), 1–14.

[43] Huang, H., Held-Feindt, J., Buhl, R., Mehdorn, H. M., and Mentlein, R. Expression of VEGF and its receptors in different brain tumors. *Neurological Research 27*, 4 (2005), 371–377.

[44] Incoronato, M., Aiello, M., Infante, T., Cavaliere, C., Grimaldi, A. M., Mirabelli, P., Monti, S., and Salvatore, M.

Radiogenomic analysis of oncological data: a technical survey. *International Journal of Molecular Sciences 18*, 4 (2017), 805.

[45] Jackson, P. R., Juliano, J., Hawkins-Daarud, A., Rockne, R. C., and Swanson, K. R. Patient-specific mathematical neuro-oncology: using a simple proliferation and invasion tumor model to inform clinical practice. *Bulletin of Mathematical Biology 77*, 5 (2015), 846–856.

[46] Jbabdi, S., Mandonnet, E., Duffau, H., Capelle, L., Swanson, K. R., Pélégrini-Issac, M., Guillevin, R., and Benali, H. Simulation of anisotropic growth of low-grade gliomas using diffusion tensor imaging. *Magnetic Resonance in Medicine 54*, 3 (2005), 616–624.

[47] Jones, D. K., and Leemans, A. Diffusion tensor imaging. *Magnetic Resonance Neuroimaging* (2011), 127–144.

[48] Keller, E. F., and Segel, L. A. Model for chemotaxis. *Journal of Theoretical Biology 30*, 2 (1971), 225–234.

[49] Kelly, P. J. Computed tomography and histologic limits in glial neoplasms: tumor types and selection for volumetric resection. *Surgical Neurology 39*, 6 (1993), 458–465.

[50] Kelly, P. J., Daumas-Duport, C., Kispert, D. B., Kall, B. A., Scheithauer, B. W., and Illig, J. J. Imaging-based stereotaxic serial biopsies in untreated intracranial glial neoplasms. *Journal of Neurosurgery 66*, 6 (1987), 865–874.

[51] Kolmogorov, A. N. A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem. *Moscow Univ. Bull. Math. 1* (1937), 1–25.

[52] Konukoglu, E., Clatz, O., Bondiau, P.-Y., Delingette, H., and Ayache, N. Extrapolating glioma invasion margin in brain magnetic resonance images: suggesting new irradiation margins. *Medical Image Analysis 14*, 2 (2010), 111–125.

[53] Lino, M., and Merlo, A. PI3Kinase signaling in glioblastoma. *Journal of Neuro-Oncology 103*, 3 (2011), 417–427.

[54] Little, S. E., Popov, S., Jury, A., Bax, D. A., Doey, L., Al-Sarraj, S., Jurgensmeier, J. M., and Jones, C. Receptor tyrosine kinase genes amplified in glioblastoma exhibit a mutual exclusivity in

variable proportions reflective of individual tumor heterogeneity. *Cancer Research 72*, 7 (2012), 1614–1620.

[55] Lopez-Gines, C., Gil-Benso, R., Ferrer-Luna, R., Benito, R., Serna, E., Gonzalez-Darder, J., Quilis, V., Monleon, D., Celda, B., and Cerdá-Nicolas, M. New pattern of EGFR amplification in glioblastoma and the relationship of gene copy number with gene expression profile. *Modern Pathology 23*, 6 (2010), 856–865.

[56] Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., Scheithauer, B. W., and Kleihues, P. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica 114*, 2 (2007), 97–109.

[57] Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., and Ellison, D. W. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica 131*, 6 (2016), 803–820.

[58] Lyons, J. G., Lobo, E., Martorana, A. M., and Myerscough, M. R. Clonal diversity in carcinomas: its implications for tumour progression and the contribution made to it by epithelial-mesenchymal transitions. *Clinical & Experimental Metastasis 25*, 6 (2008), 665–677.

[59] Madzvamuse, A., Ndakwo, H. S., and Barreira, R. Cross-diffusion-driven instability for reaction-diffusion systems: analysis and simulations. *Journal of Mathematical Biology 70*, 4 (2015), 709–743.

[60] Marino, S., Hogue, I. B., Ray, C. J., and Kirschner, D. E. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology 254*, 1 (2008), 178–196.

[61] Martirosyan, N. L., Rutter, E. M., Ramey, W. L., Kostelich, E. J., Kuang, Y., and Preul, M. C. Mathematically modeling the biological properties of gliomas: a review. *Mathematical Biosciences & Engineering 12*, 4 (2015), 879–905.

[62] Massey, S. C., Urcuyo, J. C., Marin, B. M., Sarkaria, J. N., and Swanson, K. R. Quantifying glioblastoma drug response dynamics incorporating treatment sensitivity and blood brain barrier penetrance from experimental data. *Frontiers in Physiology* (2020), 830.

[63] Massey, S. C., White, H., Rayfield, C., Rickertsen, C. R., Clark-Swanson, K., Whitmire, S., Johnston, S. K., Porter, A., Mrugala, M., Bendok, B., and Swanson, K. R. Extent of glioblastoma invasion predicts overall survival following upfront radiotherapy concurrent with temozolomide. *Neuro-Oncology Volume 19*, Suppl 6 (2017).

[64] McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statistical Science 33*, 1 (2018), 4–18.

[65] Mehrian-Shai, R., Yalon, M., Moshe, I., Barshack, I., Nass, D., Jacob, J., Dor, C., Reichardt, J. K., Constantini, S., and Toren, A. Identification of genomic aberrations in hemangioblastoma by droplet digital PCR and SNP microarray highlights novel candidate genes and pathways for pathogenesis. *BMC Genomics 17*, 1 (2016), 1–11.

[66] Minter, A., and Retkute, R. Approximate Bayesian computation for infectious disease modelling. *Epidemics 29* (2019), 100368.

[67] Molina, D., Pérez-Beteta, J., Luque, B., Arregui, E., Calvo, M., Borrás, J. M., López, C., Martino, J., Velasquez, C., Asenjo, B., et al. Tumour heterogeneity in glioblastoma assessed by MRI texture analysis: a potential marker of survival. *The British Journal of Radiology 89*, 1064 (2016), 20160242.

[68] Morris, B., Curtin, L., Hawkins-Daarud, A., Hubbard, M. E., Rahman, R., Smith, S. J., Auer, D., Tran, N. L., Hu, L. S., Eschbacher, J. M., et al. Identifying the spatial and temporal dynamics of molecularly-distinct glioblastoma sub-populations. *Mathematical Biosciences and Engineering 17*, 5 (2020), 4905–4941.

[69] Mosayebi, P., Cobzas, D., Murtha, A., and Jagersand, M. Tumor invasion margin on the Riemannian space of brain fibers. *Medical Image Analysis 16*, 2 (2012), 361–373.

[70] Müller, C., Holtschmidt, J., Auer, M., Heitzer, E., Lamszus, K., Schulte, A., Matschke, J., Langer-Freitag, S., Gasch, C., Stoupiec, M., et al. Hematogenous dissemination of glioblastoma multiforme. *Science Translational Medicine 6*, 247 (2014), 247ra101–247ra101.

[71] Murray, J. D. *Mathematical biology II: spatial models and biomedical applications*, vol. 3. Springer New York, 2001.

[72] Nakada, M., Kita, D., Watanabe, T., Hayashi, Y., and Hamada, J.-i. The mechanism of chemoresistance against tyrosine kinase inhibitors in malignant glioma. *Brain Tumor Pathology 31*, 3 (2014), 198–207.

[73] Neal, M. L., Trister, A. D., Ahn, S., Baldock, A., Bridge, C. A., Guyman, L., Lange, J., Sodt, R., Cloke, T., Lai, A., et al. Response classification based on a minimal model of glioblastoma growth is prognostic for clinical outcomes and distinguishes progression from pseudoprogression. *Cancer Research 73*, 10 (2013), 2976–2986.

[74] Neal, M. L., Trister, A. D., Cloke, T., Sodt, R., Ahn, S., Baldock, A. L., Bridge, C. A., Lai, A., Cloughesy, T. F., Mrugala, M. M., et al. Discriminating survival outcomes in patients with glioblastoma using a simulation-based, patient-specific response metric. *PloS One 8*, 1 (2013), e51951.

[75] Ostrander, S. Macroscopic cross-diffusion models derived from spatially discrete continuous time microscopic models. *SIAM Undergrad. Res. Online 4* (2011), 51–71.

[76] Painter, K., and Hillen, T. Mathematical modelling of glioma growth: the use of diffusion tensor imaging (DTI) data to predict the anisotropic pathways of cancer invasion. *Journal of Theoretical Biology 323* (2013), 25–39.

[77] Painter, K. J. Continuous models for cell migration in tissues and applications to cell sorting via differential chemotaxis. *Bulletin of Mathematical Biology 71*, 5 (2009), 1117.

[78] Painter, K. J., and Hillen, T. Volume-filling and quorum-sensing in models for chemosensitive movement. *Can. Appl. Math. Quart 10*, 4 (2002), 501–543.

[79] Papadogiorgaki, M., Koliou, P., Kotsiakis, X., and Zervakis, M. E. Mathematical modelling of spatio-temporal glioma evolution. *Theoretical Biology and Medical Modelling 10*, 1 (2013), 47.

[80] Parker, J. J., Canoll, P., Niswander, L., Kleinschmidt-DeMasters, B., Foshay, K., and Waziri, A. Intratumoral heterogeneity of endogenous tumor cell invasive behavior in human glioblastoma. *Scientific Reports 8*, 1 (2018), 1–10.

[81] Parker, J. J., Dionne, K. R., Massarwa, R., Klaassen, M., Foreman, N. K., Niswander, L., Canoll, P., Kleinschmidt-Demasters, B., and Waziri, A. Gefitinib selectively inhibits tumor cell migration in EGFR-amplified human glioblastoma. *Neuro-Oncology 15*, 8 (2013), 1048–1057.

[82] Parker, N. R., Khong, P., Parkinson, J. F., Howell, V. M., and Wheeler, H. R. Molecular heterogeneity in glioblastoma: potential clinical implications. *Frontiers in Oncology 5* (2015), 55.

[83] Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., et al. An integrated genomic analysis of human glioblastoma multiforme. *Science 321*, 5897 (2008), 1807–1812.

[84] Parvez, K., Parvez, A., and Zadeh, G. The diagnosis and treatment of pseudoprogression, radiation necrosis and brain tumor recurrence. *International Journal of Molecular Sciences 15*, 7 (2014), 11832–11846.

[85] Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* (2014), 1254257.

[86] Patel, V., and Hathout, L. Image-driven modeling of the proliferation and necrosis of glioblastoma multiforme. *Theoretical Biology and Medical Modelling 14*, 1 (2017), 10.

[87] Patlak, C. S. Random walk with persistence and external bias. *The Bulletin of Mathematical Biophysics 15*, 3 (1953), 311–338.

[88] Pérez-Beteta, J., Belmonte-Beitia, J., and Pérez-García, V. M. Tumor width on T1-weighted MRI images of glioblastoma as a prognostic biomarker: a mathematical model. *Mathematical Modelling of Natural Phenomena 15* (2020), 10.

[89] Protopappa, M., Zygogianni, A., Stamatakos, G., Antypas, C., Armpilia, C., Uzunoglu, N., and Kouloulias, V. Clinical implications of in silico mathematical modeling for glioblastoma: a critical review. *Journal of Neuro-Oncology 136*, 1 (2018), 1–11.

[90] Raizer, J. J., Abrey, L. E., Lassman, A. B., Chang, S. M., Lamborn, K. R., Kuhn, J. G., Yung, W. A., Gilbert, M. R., Aldape, K. A., Wen, P. Y., et al. A phase II trial of erlotinib in patients with recurrent malignant gliomas and nonprogressive glioblastoma multiforme postradiation therapy. *Neuro-Oncology 12*, 1 (2009), 95–103.

[91] Reardon, D. A., Fink, K. L., Mikkelsen, T., Cloughesy, T. F., O'Neill, A., Plotkin, S., Glantz, M., Ravin, P., Raizer, J. J., Rich, K. M., et al. Randomized phase II study of cilengitide, an integrin-targeting arginine-glycine-aspartic acid peptide, in recurrent glioblastoma multiforme. *Journal of Clinical Oncology 26*, 34 (2008), 5610–5617.

[92] Rees, J. Advances in magnetic resonance imaging of brain tumours. *Current Opinion in Neurology 16*, 6 (2003), 643–650.

[93] Rockne, R., Rockhill, J., Mrugala, M., Spence, A., Kalet, I., Hendrickson, K., Lai, A., Cloughesy, T., Alvord Jr, E., and Swanson, K. Predicting the efficacy of radiotherapy in individual glioblastoma patients in vivo: a mathematical modeling approach. *Physics in Medicine & Biology 55*, 12 (2010), 3271.

[94] Sánchez-Garduño, F., and Maini, P. K. Existence and uniqueness of a sharp travelling wave in degenerate non-linear diffusion Fisher-KPP equations. *Journal of Mathematical Biology 33*, 2 (1994), 163–192.

[95] Sánchez-Garduño, F., and Maini, P. K. Travelling wave phenomena in non-linear diffusion degenerate Nagumo equations. *Journal of Mathematical Biology 35*, 6 (1997), 713–728.

[96] Sherratt, J. A., and Nowak, M. A. Oncogenes, anti-oncogenes and the immune response to cancer: a mathematical model. *Proceedings of the Royal Society of London. Series B: Biological Sciences 248*, 1323 (1992), 261–271.

[97] Shinojima, N., Tada, K., Shiraishi, S., Kamiryo, T., Kochi, M., Nakamura, H., Makino, K., Saya, H., Hirano, H., Kuratsu, J.-i.,

ET AL. Prognostic value of epidermal growth factor receptor in patients with glioblastoma multiforme. *Cancer Research 63*, 20 (2003), 6962–6970.

[98] SILBERGELD, D. L., AND CHICOINE, M. R. Isolation and characterization of human malignant glioma cells from histologically normal brain. *Journal of Neurosurgery 86*, 3 (1997), 525–531.

[99] SILK, D., FILIPPI, S., AND STUMPF, M. P. Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Statistical Applications in Genetics and Molecular Biology 12*, 5 (2013), 603–618.

[100] SINGLETON, K. W., JOHNSTON, S. K., HAWKINS-DAARUD, A., RICKERTSEN, C. R., DE LEON, G., KUNKEL, L. R., WHITMIRE, S. A., CLARK-SWANSON, K., BENDOK, B., MRUGALA, M., PORTER, A., AND SWANSON, K. R. Discrimination of clinically impactful treatment response in recurrent glioblastoma patients receiving bevacizumab treatment. *Soceity of Neuro-Oncology* (2017). Abstract. In Press.

[101] SINGLETON, K. W., JOHNSTON, S. K., RICKERTSEN, C. R., DE LEON, G., KUNKEL, L. R., ROCKHILL, J., MRUGALA, M., BENDOK, B., PATEL, N., PORTER, A., AND SWANSON, K. R. Role of pre-treatment tumour dynamics and imaging response in discriminating glioblastoma survival following gamma knife. *Society of Neuro-Oncology* (2017). Abstract. In Press.

[102] SIVAKUMAR, H., DEVARASETTY, M., KRAM, D. E., STROWD, R. E., AND SKARDAL, A. Multi-cell type glioblastoma tumor spheroids for evaluating sub-population-specific drug response. *Frontiers in Bioengineering and Biotechnology 8* (2020), 1096.

[103] SKAGA, E., KULESSKIY, E., FAYZULLIN, A., SANDBERG, C. J., POTDAR, S., KYTTÄLÄ, A., LANGMOEN, I. A., LAAKSO, A., GAÁL-PAAVOLA, E., PEROLA, M., ET AL. Intertumoral heterogeneity in patient-specific drug sensitivities in treatment-naïve glioblastoma. *BMC Cancer 19*, 1 (2019), 1–14.

[104] SMITH, C. A., AND YATES, C. A. The auxiliary region method: a hybrid method for coupling PDE-and Brownian-based dynamics for reaction–diffusion systems. *Royal Society Open Science 5*, 8 (2018), 180920.

[105] Smith, S., Diksin, M., Chhaya, S., Sairam, S., Estevez-Cebrero, M., and Rahman, R. The invasive region of glioblastoma defined by 5ALA guided surgery has an altered cancer stem cell marker profile compared to central tumour. *International Journal of Molecular Sciences 18*, 11 (2017), 2452.

[106] Smith, S. J., Rahman, C. V., Clarke, P., Ritchie, A., Gould, T., Ward, J. H., Shakesheff, K. M., Grundy, R. G., and Rahman, R. Surgical delivery of drug releasing poly (lactic-co-glycolic acid)/poly (ethylene glycol) paste with in vivo effects against glioblastoma. *The Annals of The Royal College of Surgeons of England 96*, 7 (2014), 495–501.

[107] Snuderl, M., Fazlollahi, L., Le, L. P., Nitta, M., Zhelyazkova, B. H., Davidson, C. J., Akhavanfard, S., Cahill, D. P., Aldape, K. D., Betensky, R. A., et al. Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell 20*, 6 (2011), 810–817.

[108] Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C., and Tavaré, S. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences 110*, 10 (2013), 4009–4014.

[109] Stevens, A., and Othmer, H. G. Aggregation, blowup, and collapse: the ABC's of taxis in reinforced random walks. *SIAM Journal on Applied Mathematics 57*, 4 (1997), 1044–1081.

[110] Strobl, M. A., Gallaher, J., West, J., Robertson-Tessi, M., Maini, P. K., and Anderson, A. R. Spatial structure impacts adaptive therapy by shaping intra-tumoral competition. *Communications Medicine 2*, 1 (2022), 1–18.

[111] Stupp, R., Hegi, M. E., Mason, W. P., Van Den Bent, M. J., Taphoorn, M. J., Janzer, R. C., Ludwin, S. K., Allgeier, A., Fisher, B., Belanger, K., et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *The Lancet Oncology 10*, 5 (2009), 459–466.

[112] Stupp, R., Hegi, M. E., Neyns, B., Goldbrunner, R., Schlegel, U., Clement, P. M., Grabenbauer, G. G., Ochsenbein, A. F., Simon, M., Dietrich, P.-Y., et al. Phase I/IIa study of cilengitide and temozolomide with concomitant radiotherapy followed by cilengitide and temozolomide maintenance therapy in patients with newly diagnosed glioblastoma. *Journal of Clinical Oncology 28*, 16 (2010), 2712–2718.

[113] Stupp, R., Mason, W. P., Van Den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., Belanger, K., Brandes, A. A., Marosi, C., Bogdahn, U., et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine 352*, 10 (2005), 987–996.

[114] Swan, A., Hillen, T., Bowman, J. C., and Murtha, A. D. A patient-specific anisotropic diffusion model for brain tumour spread. *Bulletin of Mathematical Biology 80*, 5 (2018), 1259–1291.

[115] Swanson, K., Harpold, H., Peacock, D., Rockne, R., Pennington, C., Kilbride, L., Grant, R., Wardlaw, J., and Alvord, E. Velocity of radial expansion of contrast-enhancing gliomas and the effectiveness of radiotherapy in individual patients: a proof of principle. *Clinical Oncology 20*, 4 (2008), 301–308.

[116] Swanson, K., Rostomily, R., and Alvord Jr, E. A mathematical modelling tool for predicting survival of individual patients following resection of glioblastoma: a proof of principle. *British Journal of Cancer 98*, 1 (2008), 113.

[117] Swanson, K. R., Alvord, E., and Murray, J. A quantitative model for differential motility of gliomas in grey and white matter. *Cell Proliferation 33*, 5 (2000), 317–329.

[118] Swanson, K. R., Alvord, E. C., Murray, J. D., Rockne, R., et al. Method and system for characterizing tumors, Oct. 29 2013. US Patent 8,571,844.

[119] Swanson, K. R., Alvord Jr, E. C., and Murray, J. Virtual brain tumours (gliomas) enhance the reality of medical imaging and highlight inadequacies of current therapy. *British Journal of Cancer 86*, 1 (2002), 14.

[120] Swanson, K. R., Alvord Jr, E. C., and Rostomily, R. C. Confirmation of a theoretical model describing the relative contributions net

growth and dispersal in individual infiltrating gliomas. *Can J Neurol Sci 30* (2003), 407.

[121] Swanson, K. R., Bridge, C., Murray, J., and Alvord, E. C. Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. *Journal of the Neurological Sciences 216*, 1 (2003), 1–10.

[122] Swanson, K. R., Rockne, R. C., Claridge, J., Chaplain, M. A., Alvord, E. C., and Anderson, A. R. Quantifying the role of angiogenesis in malignant progression of gliomas: in silico modeling integrates imaging and histology. *Cancer Research 71*, 24 (2011), 7366–7375.

[123] Szerlip, N. J., Pedraza, A., Chakravarty, D., Azim, M., McGuire, J., Fang, Y., Ozawa, T., Holland, E. C., Huse, J. T., Jhanwar, S., et al. Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. *Proceedings of the National Academy of Sciences 109*, 8 (2012), 3041–3046.

[124] Talasila, K. M., Soentgerath, A., Euskirchen, P., Rosland, G. V., Wang, J., Huszthy, P. C., Prestegarden, L., Skaftnesmo, K. O., Sakariassen, P. Ø., Eskilsson, E., et al. EGFR wild-type amplification and activation promote invasion and development of glioblastoma independent of angiogenesis. *Acta Neuropathologica 125*, 5 (2013), 683–698.

[125] The Cancer Genome Atlas (TCGA) Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature 455*, 7216 (2008), 1061.

[126] Thiessen, B., Stewart, C., Tsao, M., Kamel-Reid, S., Schaiquevich, P., Mason, W., Easaw, J., Belanger, K., Forsyth, P., McIntosh, L., et al. A phase I/II trial of GW572016 (lapatinib) in recurrent glioblastoma multiforme: clinical outcomes, pharmacokinetics and molecular correlation. *Cancer Chemotherapy and Pharmacology 65*, 2 (2010), 353–361.

[127] Toni, T. *Approximate Bayesian computation for parameter inference and model selection in systems biology.* PhD thesis, Imperial College London, 2010.

[128] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface 6*, 31 (2009), 187–202.

[129] Voldborg, B. R., Damstrup, L., Spang-Thomsen, M., and Poulsen, H. S. Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials. *Annals of Oncology 8*, 12 (1997), 1197–1206.

[130] Wang, C. H., Rockhill, J. K., Mrugala, M., Peacock, D. L., Lai, A., Jusenius, K., Wardlaw, J. M., Cloughesy, T., Spence, A. M., Rockne, R., et al. Prognostic significance of growth kinetics in newly diagnosed glioblastomas revealed by combining serial imaging with a novel biomathematical model. *Cancer Research 69*, 23 (2009), 9133–9140.

[131] Wang, L., D'Angelo, F., Curtin, L., Sereduk, C. P., De Leon, G., Singleton, K. W., Urcuyo, J., Hawkins-Daarud, A., et al. Quantifying intra-tumoral genetic heterogeneity of glioblastoma toward precision medicine using MRI and a data-inclusive machine learning algorithm. *In Submission* (2022).

[132] Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., DeGroot, J., Wick, W., Gilbert, M. R., Lassman, A. B., et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *Journal of Clinical Oncology 28*, 11 (2010), 1963–1972.

[133] Woodward, D. i. w., Cook, J., Tracqui, P., Cruywagen, G., Murray, J., and Alvord, E. A mathematical model of glioma growth: the effect of extent of surgical resection. *Cell Proliferation 29*, 6 (1996), 269–288.

[134] Zhang, B., Chang, K., Ramkissoon, S., Tanguturi, S., Bi, W. L., Reardon, D. A., Ligon, K. L., Alexander, B. M., Wen, P. Y., and Huang, R. Y. Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro-Oncology 19*, 1 (2016), 109–117.

[135] Zhang, L., Strouthos, C. G., Wang, Z., and Deisboeck, T. S. Simulating brain tumor heterogeneity with a multiscale agent-based model:

linking molecular signatures, phenotypes and expansion rate. *Mathematical and Computer Modelling 49*, 1-2 (2009), 307–319.

Appendix A

Finite Difference Scheme for Interacting Species Model in 1D

We derive a finite difference scheme to solve the one-dimensional case of the model, given by Eq.s (2.1)-(2.7), over the finite domain $(x,t) \in [0,L] \times [0,T]$. We discretise our spatial domain into a mesh of $R+1$ evenly spaced points with uniform mesh-spacing h, such that $Rh = L$; similarly, we obtain a mesh of $S+1$ time points with time-step τ, such that $N\tau = T$. We introduce the notation $E_i^n = E(ih, n\tau)$ for $i = 0, ..., R$ and $n = 0, ..., S$, with P_i^n and N_i^n defined analogously.

Numerical approximation of the movement term in the model (with D_E constant): We approximate this term in 2 steps. First, the outside derivative.

$$\begin{aligned}
&\frac{\partial}{\partial x}\left(D_E\left(1-\frac{P_i^n+N_i^n}{K}\right)\frac{\partial E_i^n}{\partial x}+D_E\frac{E_i^n}{K}\left(\frac{\partial P_i^n}{\partial x}+\frac{\partial N_i^n}{\partial x}\right)\right)\\
&\quad=\frac{1}{h}\left[D_E\left(1-\frac{P_{i+1/2}^n+N_{i+1/2}^n}{K}\right)\frac{\partial E_{i+1/2}^n}{\partial x}+D_E\frac{E_{i+1/2}^n}{K}\left(\frac{\partial P_{i+1/2}^n}{\partial x}+\frac{\partial N_{i+1/2}^n}{\partial x}\right)\right.\\
&\quad\left.-D_E\left(1-\frac{P_{i-1/2}^n+N_{i-1/2}^n}{K}\right)\frac{\partial E_{i-1/2}^n}{\partial x}-D_E\frac{E_{i-1/2}^n}{K}\left(\frac{\partial P_{i-1/2}^n}{\partial x}+\frac{\partial N_{i-1/2}^n}{\partial x}\right)\right]
\end{aligned} \tag{A.1}$$

Next, approximating the remaining derivatives and the terms that lie between mesh points with, for example,

$$E_{i+1/2}^n = \frac{E_i^n + E_{i+1}^n}{2}, \tag{A.2}$$

we have

$$\begin{aligned}
&\frac{\partial}{\partial x}\left(D_E\left(1-\frac{P_i^n+N_i^n}{K}\right)\frac{\partial E_i^n}{\partial x}+D_E\frac{E_i^n}{K}\left(\frac{\partial P_i^n}{\partial x}+\frac{\partial N_i^n}{\partial x}\right)\right)\\
&\quad=\frac{D_E}{h^2}\left[\left(1-\frac{P_{i+1}^n+P_i^n+N_{i+1}^n+N_i^n}{2K}\right)(E_{i+1}^n-E_i^n)\right.\\
&\quad\quad+\left(\frac{E_{i+1}^n+E_i^n}{2K}\right)(P_{i+1}^n-P_i^n+N_{i+1}^n-N_i^n)\\
&\quad\quad-\left(1-\frac{P_i^n+P_{i-1}^n+N_i^n+N_{i-1}^n}{2K}\right)(E_i^N-E_{i-1}^n)\\
&\quad\quad\left.-\left(\frac{E_i^n+E_{i-1}^n}{2K}\right)(P_i^n-P_{i-1}^n+N_i^n-N_{i-1}^n)\right]\\
&\quad=:g_{E,i}^n,\ \text{ for }\ i=1,...,R-1\ \text{ and }\ n=0,....,S.
\end{aligned}\tag{A.3}$$

We note that the zero flux boundary conditions require us to treat the cases where $i=0$ and R with some caution. To account for these boundary conditions, we introduce ghost mesh points where $E_{-1}^n=E_1^n$ and $E_{R+1}^n=E_{R-1}^n$ for $n=0,...,S$, with the P and N species cases analogously defined. Therefore, we have

$$\begin{aligned}
g_{E,0}^n:=\frac{2D_E}{h^2}&\left[\left(1-\frac{P_1^n+P_0^n+N_1^n+N_0^n}{2K}\right)(E_1^n-E_0^n)\right.\\
&\left.+\left(\frac{E_1^n+E_0^n}{2K}\right)(P_1^n-P_0^n+N_1^n-N_0^n)\right],
\end{aligned}\tag{A.4}$$

$$\begin{aligned}
g_{E,R}^n:=\frac{2D_E}{h^2}&\left[\left(1-\frac{P_{R-1}^n+P_R^n+N_{R-1}^n+N_R^n}{2K}\right)(E_{R-1}^n-E_R^n)\right.\\
&\left.+\left(\frac{E_{R-1}^n+E_R^n}{2K}\right)(P_{R-1}^n-P_R^n+N_{R-1}^n-N_R^n)\right],
\end{aligned}\tag{A.5}$$

for $n=0,...,S$. Defining,

$$f_{E,i}^n=f_E(E(ih,n\tau),P(ih,n\tau),N(ih,n\tau)),\tag{A.6}$$

where f_E is defined by Eq. (2.5) and using a forward Euler time-stepping scheme, we have

$$E_i^{n+1}=\tau(g_{E,i}^n+f_{E,i}^n)+E_i^n\tag{A.7}$$

for the E population of tumour cells. The numerical schemes to simulate the P and N populations are derived analogously.

Appendix B

Validation of travelling wave speeds

In Section 2.3.2, we claim that the travelling waves that emerge in simulations of the model shown in Fig.s 2.5(a), (d) and (f) propagate with speed $2\sqrt{\rho_E D_E}$. Figure B.1 shows plots of the speed of the propagating wave in each of these simulations. From these, we observe that after an initial transition period during which the wavefront forms, the front then propagates with a speed that tends towards $2\sqrt{\rho_E D_E}$ in each case.

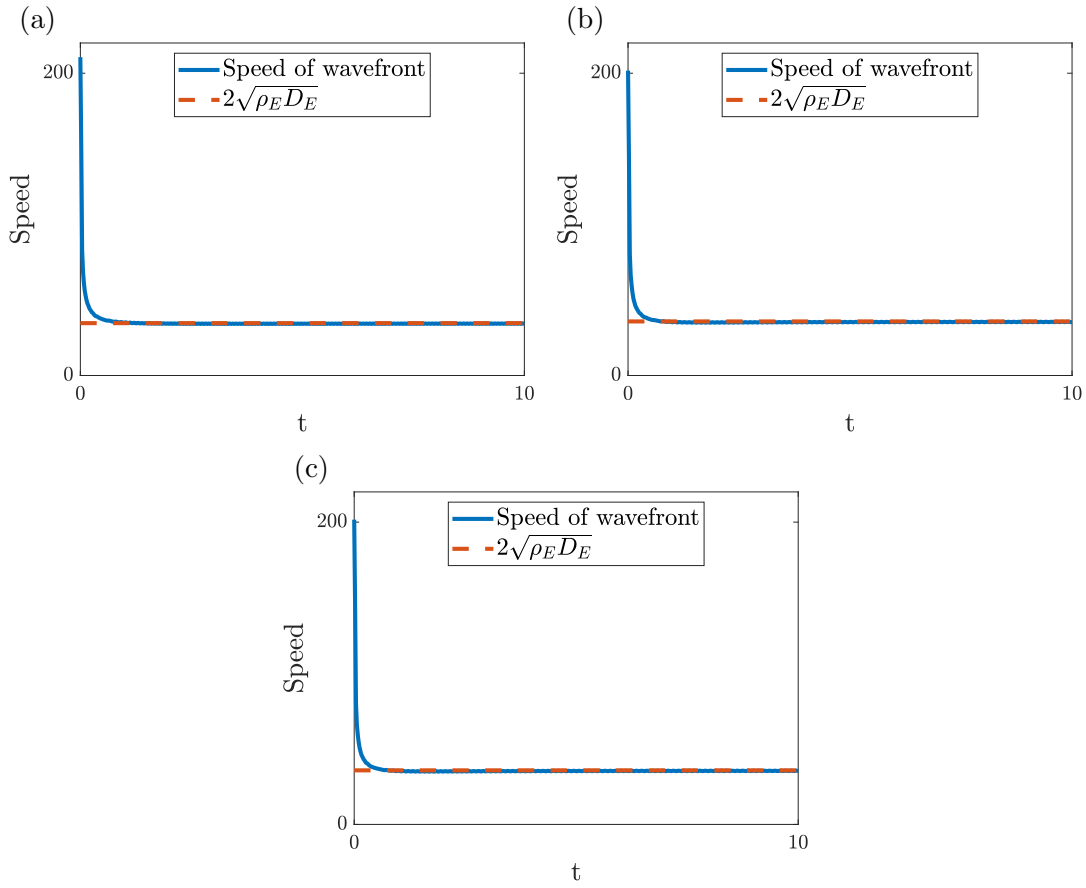


Figure B.1: Plots showing the speeds of the travelling waves shown in (a) Fig. 2.5a, (b) Fig. 2.5d and (c) Fig. 2.5f. Simulations of the model were produced under different example parameter regimes on a one-dimensional domain, $x \in [0, 400]$ mm, and over a time domain of $t \in [0, 10]$ years. In each case $D_E = D_P = D_N = 20$ mm^{2}year^{-1} and $\rho_P = \rho_N = 15$ years^{-1}, while the interaction types and proliferative ability of the E population were varied: (a) $\alpha_{EP} = \alpha_{PE} = 0$ (neutralism) and $\rho_E = 15$ years^{-1}; (b) $\alpha_{EP} = \alpha_{PE} = 0$ and $\rho_E = 16$ years^{-1} (population E has a proliferative advantage over P and N cells); (c) $\alpha_{EP} = -5$, $\alpha_{PE} = -2$ and $\rho_E = 16$ years^{-1}. The initial conditions used in model simulations are given by Eq. (2.45). From these simulations the speed of the propagating wavefronts were calculated using linear interpolation to find the point at which the wavefront passes through a threshold of $0.5K$ at each time point. The wavespeed was then calculated as the distance travelled divided by time.

Appendix C

Additional results exploring the factors affecting amplification patterns observed in simulations

Here we present additional figures to supplement those presented in Section 3.4.2 exploring the effects of various selection advantages and the timing and positioning of EGFR and PDGFRA amplified sub-population introductions on the amplification patterns we see in our simulated tumours. These figures show the general trends in changes to the proportions of simulations with neither gene, only the EGFR gene, only the PDGFRA gene and both genes amplified that changing each of these factors produces. We note that the effects of changing each of these factors are symmetric with respect to the proportions of simulations with only EGFR and only PDGFRA amplified. For example, affording E cells a 50% proliferative advantage and P cells no advantages, produces the same simulation proportions of neither and both amplified cells as giving the P population this advantage and E no advantage, while the proportions with only one gene amplified are reflected.

All of the following figures (Figures C.1–C.5) are produced from simulations with the same parameters and assumptions outlined in Section 3.4.2, apart from where parameter differences are indicated in the figure captions.

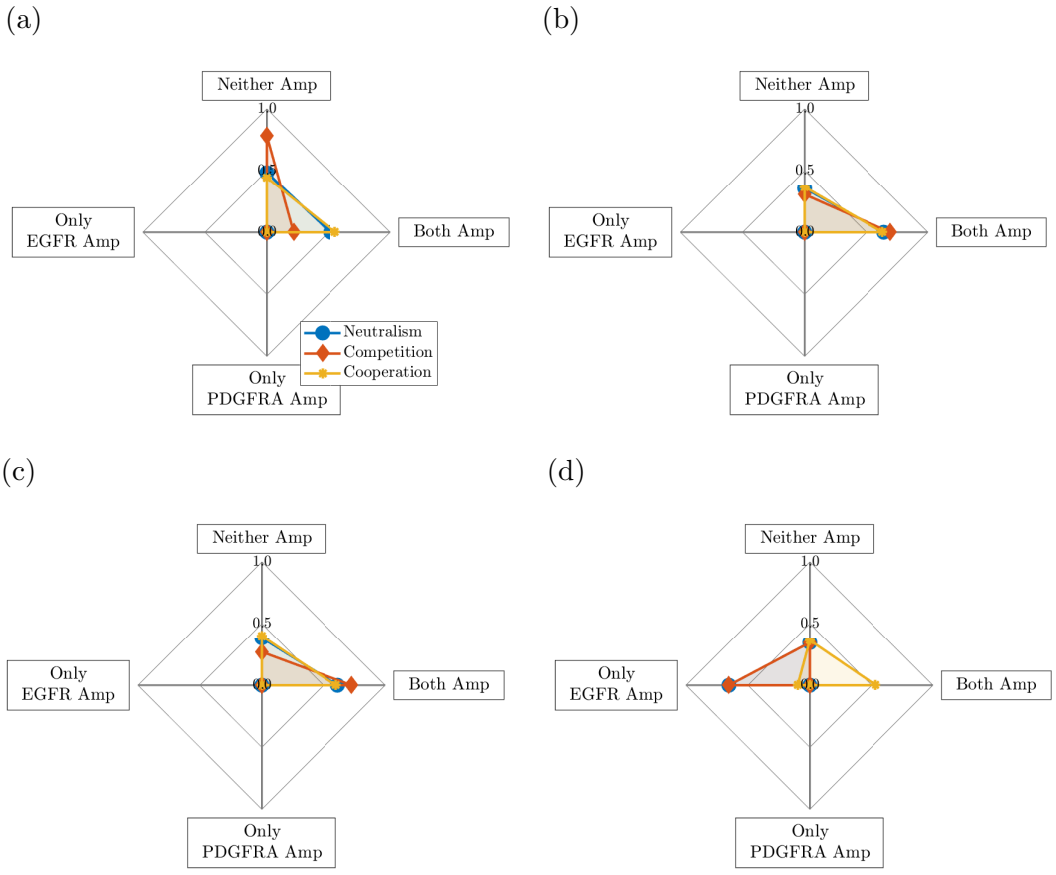


Figure C.1: Amplification patterns change when EGFR and PDGFRA amplified sub-populations are afforded proliferative and invasive advantages; proliferative advantages have the bigger impact, decreasing the proportion of the tumour with neither gene amplified. Plot showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified under different interactions when the E and P sub-populations are given various selection advantages: (a) EGFR 50% invasive advantage, PDGFRA 50% invasive advantage ($\rho_E = \rho_N$, $\rho_P = \rho_N$, $D_E = 1.5D_N$ and $D_P = 1.5D_N$); (b) EGFR 50% proliferative advantage, PDGFRA 50% proliferative advantage ($\rho_E = 1.5\rho_N$, $\rho_P = 1.5\rho_N$, $D_E = D_N$ and $D_P = D_N$); (c) EGFR 50% proliferative and invasive advantage, PDGFRA 50% proliferative and invasive advantage ($\rho_E = 1.5\rho_N$, $\rho_P = 1.5\rho_N$, $D_E = 1.5D_N$ and $D_P = 1.5D_N$); (d) EGFR 50% proliferative advantage, PDGFRA 50% invasive advantage ($\rho_E = 1.5\rho_N$, $\rho_P = \rho_N$, $D_E = D_N$ and $D_P = 1.5D_N$).

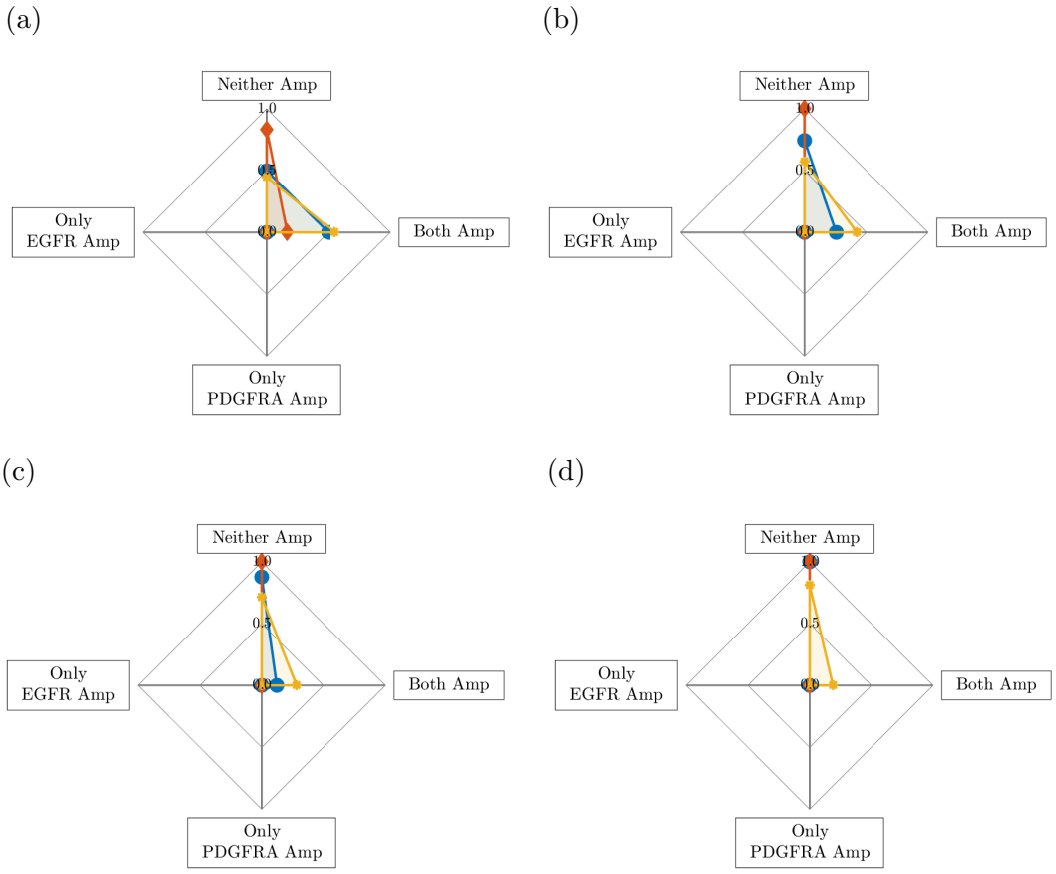


Figure C.2: Delaying the introduction of amplified sub-populations increases the proportion of the tumour with neither gene amplified and decreases the amplified proportion. Plot showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified under different interactions when the E and P sub-populations are introduced at the same time which changes: (a) $t_E^* = t_P^* = t_1^*$; (b) $t_E^* = t_P^* = t_3^*$; (c) $t_E^* = t_P^* = t_5^*$; (d) $t_E^* = t_P^* = t_7^*$, as defined in Section 3.4.2.

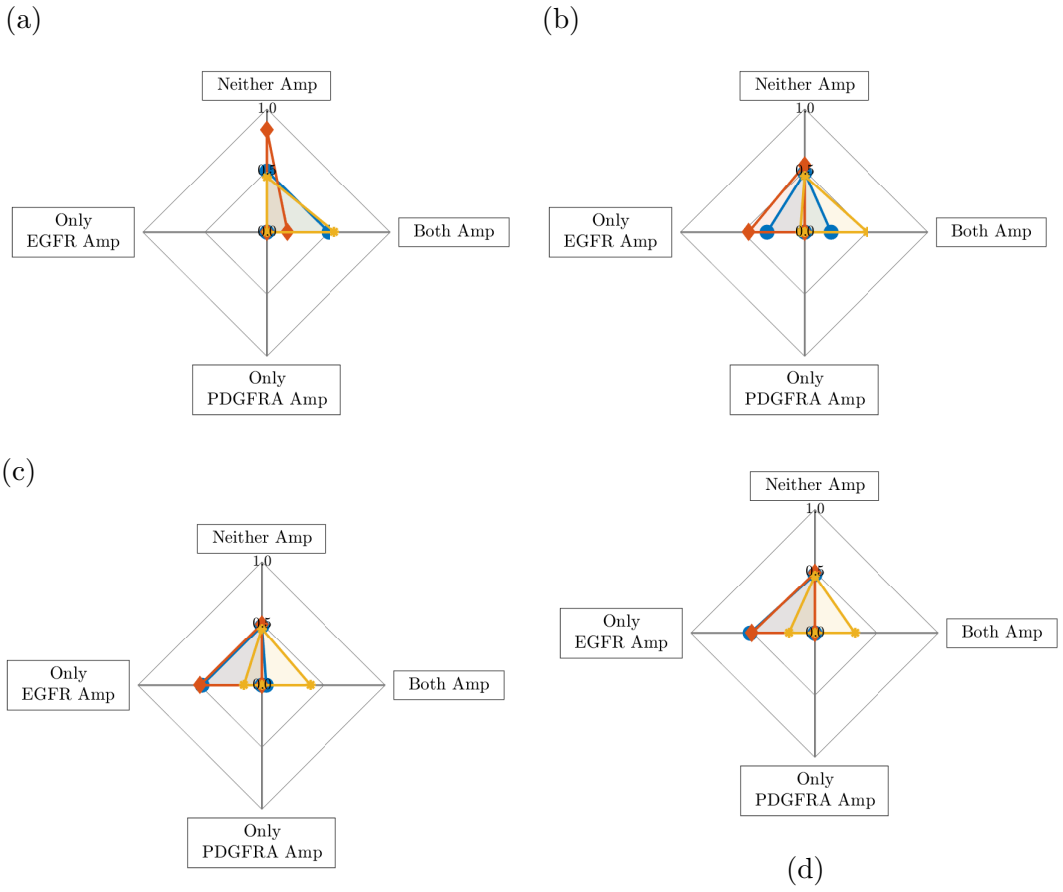


Figure C.3: Delaying the introduction of the PDGFRA amplified sub-population increases the proportion of tumour with only EGFR amplified and decreases the both amplified proportion. Plot showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified under different interactions when the E population is introduced at $t_E^* = t_1^*$ and P is introduced at: (a) $t_P^* = t_1^*$; (b) $t_P^* = t_3^*$; (c) $t_P^* = t_5^*$; (d) $t_P^* = t_7^*$, as defined in Section 3.4.2.

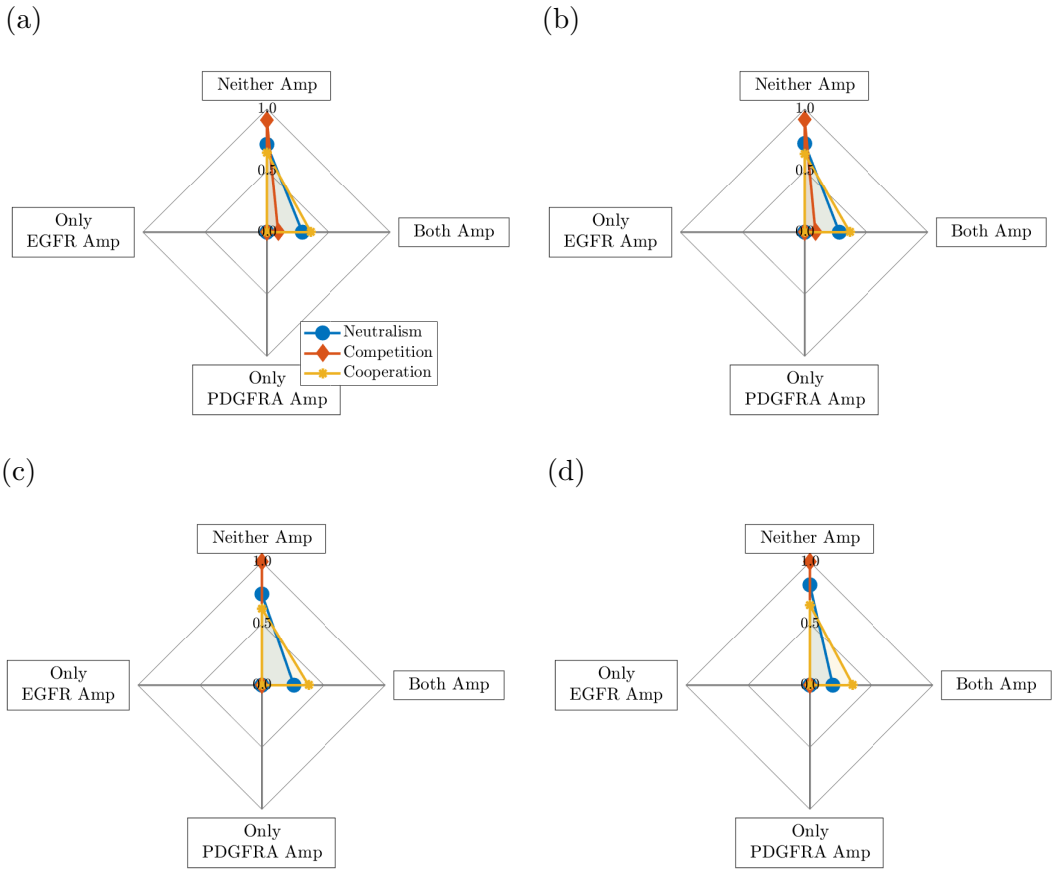


Figure C.4: Introducing amplified populations closer to the tumour centre decreases the proportion of tumour with both genes amplified in the neutral and competitive cases, although the effect is small. Plot showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified under different interactions when the E and P populations are introduced at the same location, which changes: (a) $x_E^* = x_P^* = x_1^*$; (b) $x_E^* = x_P^* = x_2^*$; (c) $x_E^* = x_P^* = x_3^*$; (d) $x_E^* = x_P^* = x_4^*$, as defined in Section 3.4.2.

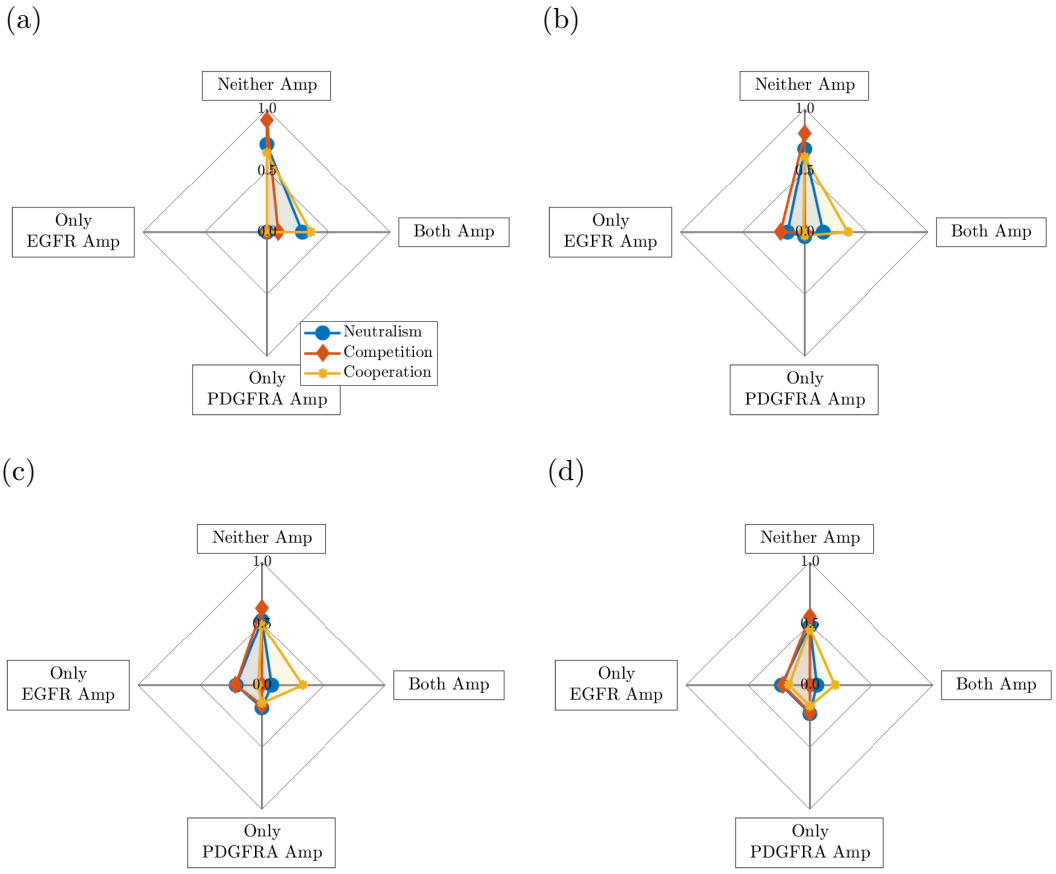


Figure C.5: Introducing PDGFRA amplified cells further from the EGFR amplified population decreases the proportion of tumour with both genes amplified and increases the proportions with only one gene amplified. Plot showing the mean proportions of simulations with neither gene (Neither Amp), only the EGFR gene (Only EGFR Amp), only the PDGFRA gene (Only PDGFRA Amp) and both genes (Both Amp) amplified under different interactions when the E population is introduced at $x_E^* = x_1^*$ and P is introduced at: (a) $x_P^* = x_1^*$; (b) $x_P^* = x_3^*$; (c) $x_P^* = x_5^*$; (d) $x_P^* = x_7^*$, as defined in Section 3.4.2.

Appendix D

Effect of non-monotonicity of introduction locations in the LHS-PRCC sensitivity analysis

Partial Rank Correlation Coefficient (PRCC) values provide a measure of the degree of monotonicity between an input and output variable and, therefore, are a good measure of sensitivity for inputs and outputs with monotonic relationships [10, 60]. As Marino et al. [60] demonstrate, a LHS-PRCC sensitivity analysis is not always accurate for input parameters and outputs with non-monotonic relationships and should be treated with caution. We observe that, due to the symmetry of our initial condition $N(x, 0)$, given by Eq. (2.4), and, thus, the growing tumour, the relationship between the introduction locations and the proportions observed in our simulations in non-monotonic. For example, in Figure C.4 choosing introduction locations $x_E^* = x_P^* = x_5^*$, x_6^* and x_7^* (as defined in Section 3.4.2) will produce equivalent results to those seen when $x_E^* = x_P^* = x_3^*$, x_2^* and x_1^*, respectively, thus non-montonically affecting the proportions of simulations with neither gene, only the EGFR gene and only the PDGFRA gene amplified. Thus, we divide the x_E^* and x_P^* domains into two, over which the relationships are monotonic and conduct a sensitivity analysis in each instance. First we study the case where both E and P are introduced on the right side of the growing tumour and we choose the uniform pdf to have minimum and maximum values of x_c^* and $x_c^* + 1.5$mm. We note that if we were to instead choose both introduction locations on the left side of the tumour, this would produce analogous results due to symmetry. Secondly, we conduct a sensitivity analysis where E and P are introduced on opposite sides of the tumour; x_E^* is selected from the right side of the tumour and x_P^* from the left. Thus, the pdf for x_E^* remains the same, but x_P^*

is instead assigned a uniform distribution with minimum and maximum values of $x_c^* - 1.5$ and x_c^*mm, respectively. All other parameters and their distributions are kept the same as detailed in Table 3.1. The results from these two LHS-PRCC sensitivity analyses are shown in Figures D.1 and D.2. In Figure D.1, we see that the PRCCs between the introduction location parameters and each of the outputs of interest do not show a strong correlation when both x_E^* and x_P^* are selected from the same side of the growing tumour. However, in Figure D.2, a weak, but significant, correlation is present between the location parameters and the proportion of simulations with only the EGFR or only the PDGFRA gene amplified when x_E^* and x_P^* are selected from opposite sides of the growing tumour, whereas there are no correlations with the other two outputs of interest. Importantly, the results for the 10 other model parameters are consistent with the main LHS-PRCC analysis presented in Section 3.4.3.

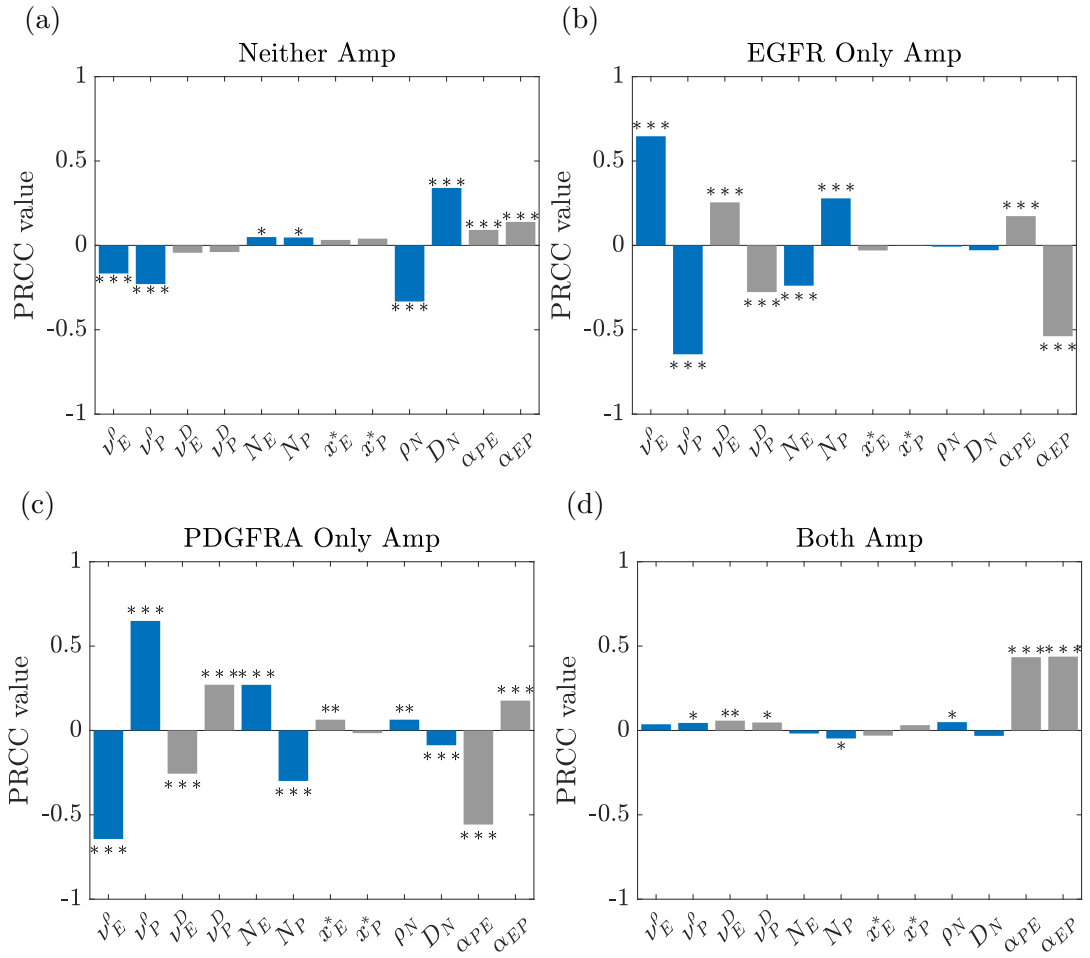


Figure D.1: Sensitivity analysis for tumour composition when both mutations arise on the same side of the tumour. Bar plots showing PRCC values between each unknown model parameter and the four outputs of interest: the proportion of simulations with (a) neither, (b) only EGFR, (c) only PDGFRA and (d) both genes amplified. All samples for the LHS step are drawn from the parameter distributions in Table 3.1, apart from the locations parameters x^*_E and x^*_P which are both drawn from a uniform distribution with minimum and maximum values of x^*_c and $x^*_c + 1.5$mm. Significant results at the 0.05 (*), the 0.01 (**) and the 0.001 (***) levels are highlighted.

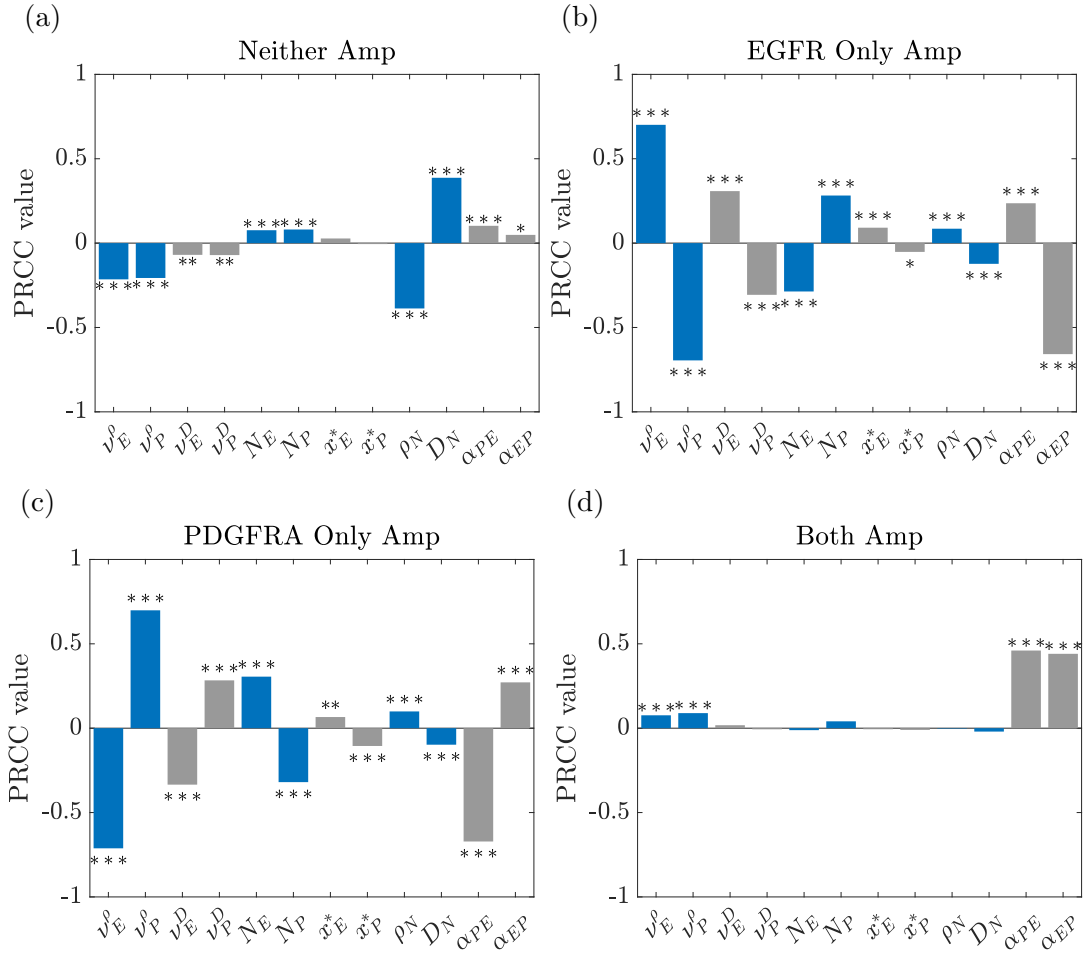


Figure D.2: Sensitivity analysis for tumour composition when the two mutations arise on opposite side of the tumour. Bar plots showing PRCC values between each unknown model parameter and the four outputs of interest: the proportion of simulations with (a) neither, (b) only EGFR, (c) only PDGFRA and (d) both genes amplified. All samples for the LHS step are drawn from the parameter distributions in Table 3.1, apart from the locations parameters x^*_E and x^*_P; x^*_E is drawn from a uniform distribution with minimum and maximum values of x^*_c and $x^*_c + 1.5$mm; x^*_P is drawn from a uniform distribution with minimum and maximum values of $x^*_c - 1.5$ and x^*_cmm. Significant results at the 0.05 (*), the 0.01 (**) and the 0.001 (***) levels are highlighted.

Appendix E

Determining the vector of tolerances for the ABC-SMC algorithm

For each implementation of the ABC-SMC algorithm in Chapter 4 we use the following approach to determine the vector of tolerances, $\epsilon = (\epsilon_1, ..., \epsilon_Q)$, where Q is the number of populations and $\epsilon_i = (\epsilon_i^M, \epsilon_i^{SD})$, for $i = 1, ..., Q$. In this work, we use the Euclidean distance as the function $d(y^*, y_)$ to determine the distance of the simulated data, y^*, from the data points, y_0. The 8 data points we use are the four mean proportions of simulations/biopsies that are amplified in neither gene, only EGFR, only PDGFRA and both genes in the simulated/patient data - we denote these four points as y_M^* and y_0^M, respectively - and the standard deviations of these proportions in the simulated/patient data - we denote as y_{SD}^* and y_0^{SD}. Rather than combine these all in one metric, termed a union metric [64], and calculate a single Euclidean distance, we choose to calculate two Euclidean distances; one for the distance between the means of the amplified proportions of the simulated and patient data and one for the distance between standard deviations. A particle is then accepted only if each of these distances are within certain tolerances; this is termed an intersection metric by McKinley et al. [64]. In this way, a particle is accepted into population i in the ABC-SMC algorithm (line 19 of the algorithm in Section 4.4.1) if

$$d(y_M^*, y_0^M) < \epsilon_i^M \quad \text{and} \quad d(y_{SD}^*, y_0^{SD}) < \epsilon_i^{SD}, \tag{E.1}$$

for $i = 1, ..., Q$.

Rather than predefine these tolerances we follow the approach of McKinley

et al. [64] and adaptively choose the values of $(\epsilon_i^M, \epsilon_i^{SD})$ for $i = 1, ..., Q$ based on the distances of the accepted particles in population $i-1$. Thus, the vector of tolerances are chosen in the following way:

After the initial population of particles have been sampled from the prior distributions and the data simulated, we choose initial tolerance values $(\epsilon_1^M, \epsilon_1^{SD})$ to be the 50th percentile of the simulated metric distances for each of the two outputs, that is the Euclidean distance between the means and standard deviations of the simulated and patient data. For $i = 1, ..., Q-1$, tolerances at generation $i+1$ are then generated using the using a bisection method (detailed in Supplement A of [64]), where the proportion of generation i particles that would be accepted using the new tolerances is approximately $p = 0.5$, where p is the target acceptance rate. We set upper and lower bounds for the target acceptance rate to be $p^U = 0.55$ and $p^L = 0.45$, respectively. The algorithm for determining $(\epsilon_i^M, \epsilon_i^{SD})$ for $i = 2, ...Q$ then proceeds as follows:

1: **for** $i = 1, ..., Q-1$ **do**
2: Run simulations with particles sampled and perturbed according to the ABC-SMC algorithm in Section 4.4.1. Accept particles into population i if they satisfy condition E.1.
3: Set the lower bounds for ϵ_{i+1}^M and ϵ_{i+1}^{SD} as the 50th percentile of the simulated distances in population i. Denote these as ϵ_L^M and ϵ_L^{SD}, respectively.
4: Set the upper bounds for ϵ_{i+1}^M and ϵ_{i+1}^{SD} as $\epsilon_U^M = \epsilon_i^M$ and $\epsilon_U^{SD} = \epsilon_i^{SD}$, respectively.
5: Set the initial proposal for ϵ_{i+1}^M and ϵ_{i+1}^{SD} as $\epsilon_*^M = \epsilon_L^M$ and $\epsilon_*^{SD} = \epsilon_L^{SD}$.
6: Calculate the proportion, p^*, of population i of accepted particles that satisfy

$$d(y_M^*, y_0^M) < \epsilon_*^M \quad \text{and} \quad d(y_{SD}^*, y_0^{SD}) < \epsilon_*^{SD}.$$

7: **if** $p^L < p^* < p^U$ **then**
8: Set $\epsilon_{i+1}^M = \epsilon_*^M$ and $\epsilon_{i+1}^{SD} = \epsilon_*^{SD}$ and stop.
9: **else if** $p^* < p^L$ **then**
10: Set $\epsilon_L^M = \epsilon_*^M$ and $\epsilon_*^M = (\epsilon_U^M + \epsilon_*^M)/2$.
11: Set $\epsilon_L^{SD} = \epsilon_*^{SD}$ and $\epsilon_*^{SD} = (\epsilon_U^{SD} + \epsilon_*^{SD})/2$.
12: Return to Step 6:
13: **else if** $p^* > p^U$ **then**
14: Set $\epsilon_U^M = \epsilon_*^M$ and $\epsilon_*^M = (\epsilon_L^M + \epsilon_*^M)/2$.
15: Set $\epsilon_U^{SD} = \epsilon_*^{SD}$ and $\epsilon_*^{SD} = (\epsilon_L^{SD} + \epsilon_*^{SD})/2$.
16: Return to Step 6:
17: **end if**

18: **end for**