Parameter identifiability analysis for spatiotemporal models of cell invasion

Matthew J Simpson[a,*], Ruth E Baker[b], Sean T Vittadello[a], Oliver J Maclaren[c]

[a] *School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia.*
[b] *Mathematical Institute, University of Oxford, Oxford, UK.*
[c] *Department of Engineering Science, University of Auckland, Auckland 1142, New Zealand.*

Abstract

We examine the practical identifiability of parameters in a spatiotemporal reaction-diffusion model of a scratch assay where we have access to a relatively large amount of quantitative experimental data. The experiment involves fluorescent cell cycle labels so that the cell cycle is tracked in real time. This provides spatial information about the position of individual cells as well as temporal information about the cell cycle phase of each cell. Cell cycle labelling is incorporated into the reaction-diffusion model by considering the total population to be composed of two interacting subpopulations. Practical identifiability is examined using a Bayesian Markov Chain Monte Carlo (MCMC) framework, confirming that the parameters are identifiable when we assume that the diffusivities of the two subpopulations are identical, but that the parameters are practically non-identifiable when we allow the diffusivities to be distinct. We also assess identifiability using a profile likelihood approach, and show that this provides similar results to MCMC with the advantage of being an order of magnitude faster to compute. Despite the profile likelihood being relatively underutilised in the mathematical biology literature, we suggest that it ought to be adopted as a screening tool to provide a rapid assessment of practical identifiability before more computationally cumbersome MCMC computations are attempted.

Keywords: Identifiability analysis; Bayesian inference; Profile likelihood; Reaction-diffusion; Cell cycle.

---

[*] Corresponding author: matthew.simpson@qut.edu.au

1. Introduction

Combined cell migration and cell proliferation leads to moving fronts of cells, often called *cell invasion* [1], which is essential for tissue repair and wound healing [2]. While sophisticated experimental techniques are continually developed to interrogate cell invasion, traditional *scratch assays* remain widely used because they are simple, fast and inexpensive [3, 4]. Scratch assays involve growing a spatially uniform monolayer of cells (Figure 1a-b), scratching the monolayer (Figure 1c), and imaging a small region around the scratch over time as the wounded area closes (Figure 1d-e).

The first mathematical models to interpret experimental observations of cell invasion were reaction-diffusion models [5]. Sherratt and Murray [6] were the first to model a wound-healing-type assay using the Fisher-Kolmogorov model and an extension of the Fisher-Kolmogorov model, called the Porous Fisher model [5]. Many subsequent studies have attempted to relate solutions of mathematical models to experimental observations of cell invasion including the work of Maini et al. [2], Sengers et al., [7], Nardini et al. [8] and Bodabilla et al. [9]. Each of these studies aim to match the solution of a mathematical model with experimental observations. This process provides estimates of model parameters which, in turn, provides mechanistic insight.

One challenge of using scratch assays is that there is no standard, widely accepted experimental protocol. There are many different experimental protocols, with differences in: (i) the initial density of the monolayer [10]; (ii) the width and shape of the scratch [1, 11]; and (iii) the time scales over which the experiment is performed [3, 4]. These differences mean that it is difficult to compare the outcomes of different experiments, and it is unclear whether different mathematical models of varying complexity (i.e. varying numbers of unknown parameters or varying the model structure) are required to describe observations made under different experimental protocols.

It is reasonable to argue that simple experimental data ought to be modelled using simple mathematical models with a small number of unknown parameters, and that more sophisticated mathematical models with additional parameters should be used only when greater amounts of experimental data are available. However, it is difficult to use this kind of qualitative reasoning to justify particular decisions about how complicated a mathematical model ought to be. Methods of *identifiability analysis* [12–20] provide more systematic means for helping to determine the appropriate model complexity relative to the data available. In particular, a model is said to be *identifiable* when it is possible, in principle, to uniquely determine the model parameters using an arbitrary amount of ideal data from a given experiment [23, 24]. This determines whether the experiment can, in principle, falsify or corroborate scientific hypotheses framed in terms of values of the model parameters. In contrast to the predictive performance focus of measures like Akaike's Information Criterion (AIC) [25], identifiability analysis is focused on distinguishing parameter values and using this to understand underlying mechanisms. Shmueli characterises this general difference in goals elsewhere as the choice of whether to 'explain or predict' [26]. In addition to identifiability in principle, the terms *practical identifiability* [17] and *estimability* [27] have been used to describe whether it is possible to provide reasonably precise and/or stable estimates using real, finite data [see also 24]. These are somewhat looser terms, as they depend on investigator

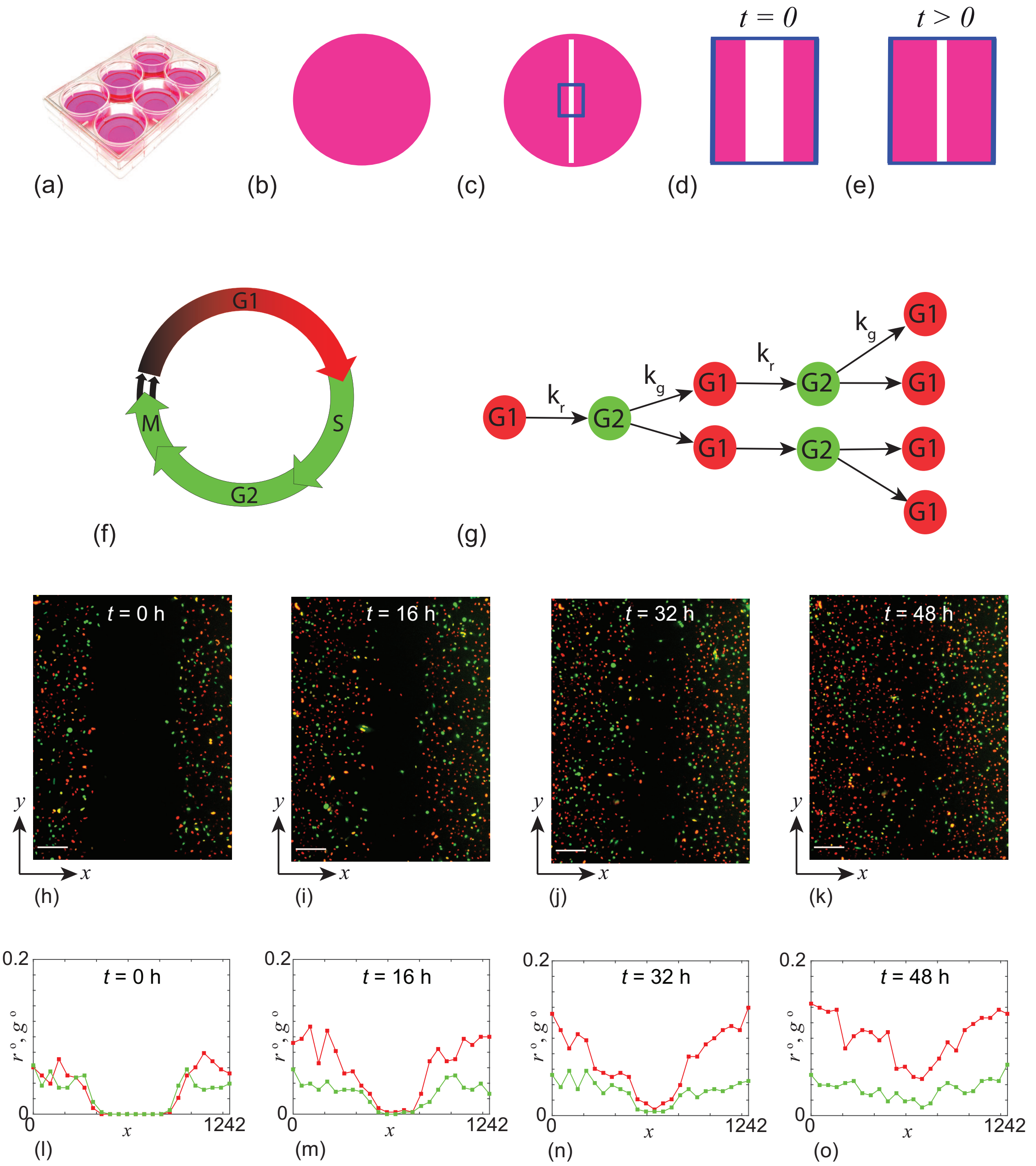

Figure 1: Experiments are performed in a 6-well plate (a), initiated with an approximately uniform monolayer of cells (b). A scratch into the monolayer is made (c), and the 1296 μm × 1745 μm field of view is observed between t = 0 h (d) and t = 48 h (e). The cell cycle is labelled using fluorescent ubiquitination-based cell cycle indicator (FUCCI) [21, 22] so that cells in G1 phase fluoresce red and cells in S/G2/M phase fluoresce green (f). A freely-cycling cell in G1 phase (red) will transition into S/G2/M phase (green) at a rate $k_r > 0$. A freely-cycling cell in S/G2/M phase (green) will divide into two cells in G1 phase (red) at rate $k_g > 0$, provided there is sufficient space to accommodate the division (g). Images of the scratch assay show the field of view at $t = 0, 16, 32$ and 48 h (h–k), with the nondimensional densities of G1 (red) and S/G2/M (green) subpopulations shown as a function of space and time (l–o). The scale bars in (l–o) show 200 μm.

goals and resources, but are usually clear in practice.

Practical identifiability analysis is closely related to power analysis in classical statistics, though the latter is based on a hypothesis-testing framework that can encourage dichotomous thinking [28–30]. Instead, identifiability analysis, in general, may be considered as encompassing a set of alternative, *precision-based* planning (pre-data) and analysis (post-data) tools similar to those advocated by Cox [28], Bland [31] and Rothman and Greenland [30]. Such methods of model identifiability analysis are well-established in the field of systems biology [12–20]. In this context experimental data often take the form of time series describing temporal variations of different biochemical molecules in some kind of chemical reaction network or gene regulatory network and these data are modelled using ordinary or stochastic differential equation models [32]. In the present work we focus on apply methods of identifiability analysis to spatiotemporal partial differential equation models, reaction-diffusion equations in particular, which model both spatial and temporal variations in different quantities of interest. We consider both a Bayesian inference framework [15, 16] and profile likelihood analysis [17, 18].

Bayesian inference is by now relatively familiar and well-adopted within the mathematical biology community [33–38]. A drawback of the Bayesian approach is that it can be computationally expensive, particularly when Markov Chain Monte Carlo (MCMC) methods are used to sample the distributions of interest. In contrast, profile likelihood analysis is a standard tool of statistical inference [28, 39], but is less familiar within the mathematical biology community. Being optimisation-based, profile likelihood can be comparatively computationally inexpensive. Furthermore, it has been argued that profile likelihood is more reliable in assessing the presence of non-identifiability than other methods, including both MCMC-based Bayesian inference [40] and Fisher-information-based approaches [18].

In this work we present images and data from a scratch assay where cells are labelled using two fluorescent probes to show real-time progression through the cell cycle [21]. Such cell cycle labelling allows us to describe the total population of cells as two subpopulations according to the cell cycle labels [41]. This experimental approach provides far more information than a standard scratch assay that does not reveal any information about the cell cycle. We model the experiments using a system of reaction-diffusion equations that are an extension of the Fisher-Kolmogorov model. Using this data and this model, we systematically explore the practical identifiability of the model parameters using both Bayesian inference and profile likelihood.

One of the main outcomes of this work is that, instead of performing potentially computationally expensive Bayesian inference to assess identifiability and determine parameter estimates, we show that a simpler profile likelihood analysis provides a fast and reliable preliminary assessment of practical identifiability. Both analysis methods provide good insight into the identifiability of the spatiotemporal models considered here. For example, we consider a minimal model scratch assay data generated using fluorescent cell cycle labelling and show that it is practically identifiable, whereas a simple extension of that model is not, despite being identifiable in principle. In both cases, the Bayesian analysis is consistent with profile likelihood results; however, the profile likelihood analysis is an order of magnitude faster to implement on a standard laptop machine.

2. Experimental data

We consider a scratch assay performed with the 1205Lu melanoma cell line with FUCCI [22]. The scratch assay is initiated by seeding 1205Lu melanoma cells into 6-well plates and allowing sufficient time for the cells to attach, begin to proliferate (Figure 1a), and to form an approximately uniform monolayer (Figure 1b). A scratch is made in the monolayer (Figure 1c) and a small region of the experiment is imaged (Figure 1c-d) [41]. The geometry of the experiment means that the distribution of cells in the imaged region is independent of vertical position, and so we treat the cell density as a function of horizontal position, x, and time, t [41]. It is important to note that the boundaries around the imaged region are not physical boundaries and cells are free to migrate across these boundaries. Since the density of cells is approximately constant away from the scratch, the net flux of cells locally across each boundary of the imaged region is approximately zero; the number of cells per unit time that move across the boundary in the positive x-direction will be approximately equal to the number of cells that move across the boundary in the negative x-direction [41].

In these experiments, cells in G1 phase fluoresce red and cells in S/G2/M phase fluoresce green (Figure 1f). We represent FUCCI labelling by treating the total population as two subpopulations: (i) red cells in G1, and (ii) green cells in S/G2/M. Conceptually, cells in G1 (red) transition into S/G2/M cells (green) at rate $k_r > 0$. This red-to-green transition does not involve any change in cell number so we assume this transition is unaffected by the availability of physical space. Cells in S/G2/M (green) undergo mitosis at rate $k_g > 0$ to produce two daughter cells in G1 (red) (Figure 1f). Since the green-to-red transition involves the production of two daughter cells we assume this transition only occurs provided there is sufficient space.

Images of the experiment (Figure 1h-k) show that cells move into the initially scratched region as well as progressing through the cell cycle. At the beginning of the experiment the scratch is approximately 500 μm wide, and by the end of the experiment the initially-vacant wound space is occupied as cells have migrated into the initially-scratched region. Visual examination of the density of cells away from the scratched region shows that the cell density increases and this increase is driven by cell proliferation. To quantify these visual observations we note that each experimental image is 1296 μm wide, and we discretise each image into 24 equally-spaced columns of width 54 μm. We count the number of red cells and the number of green cells per column, divide by the area of the column and by the theoretical maximum packing density of 0.004 cells/μm^2 [41] to give an estimate of the nondimensional density of red cells and green cells per column. Assuming that each experimental density estimate represents the centre of each column, our density estimates represent cell densities at positions $x_i = 54i$, $i = 0, 1, 2, \ldots, 23$. Plotting these data (Figure 1l–o) shows the spatial variation in density of G1 (red) and S/G2/M (green) cells as a function of space and time.

We summarise the experimental data using the following notation. Our observations are vectors containing the observed nondimensional density of red cells, r_i^{o}, and the observed nondimensional density of green cells, g_i^{o}, per column at each position x_i and time t_i, i.e.

$$(r^{\text{o}}, g^{\text{o}}, x, t)_i = (r_i^{\text{o}}, g_i^{\text{o}}, x_i, t_i). \tag{1}$$

We use an explicit superscript ‘o’ to distinguish (noisy) observations of r and g from their modelled counterparts introduced below, while we assume that the space and time coordinates x and t, respectively, are noise-free. In particular we use

$$y_i^{\mathrm{o}} = (r^{\mathrm{o}}, g^{\mathrm{o}})_i \quad \text{and} \quad s_i = (x_i, t_i), \tag{2}$$

to denote the ith combined vector of observations of r and g, and the ith combined vector of spatiotemporal coordinates x and t, respectively. Observations are taken at 24 equally-spaced spatial locations, where the spacing between the observations is 54 μm, at each of four equally-spaced time points, $t = 0$, 16, 32 and 48 h. Thus y^{o} consists of $N = 24 \times 4 = 96$ observations of both of r and g.

3. Mathematical model

A reasonably simple model of the FUCCI scratch assay can be written as

$$\frac{\partial r(x,t)}{\partial t} = D_r \frac{\partial^2 r(x,t)}{\partial x^2} - k_r r(x,t) + 2k_g g(x,t)\left[1 - r(x,t) - g(x,t)\right], \tag{3}$$

$$\frac{\partial g(x,t)}{\partial t} = D_g \frac{\partial^2 g(x,t)}{\partial x^2} + k_r r(x,t) - k_g g(x,t)\left[1 - r(x,t) - g(x,t)\right], \tag{4}$$

where $D_r > 0$ is the diffusivity of cells in G1, $D_g > 0$ is the diffusivity of cells in S/G2/M, $k_r > 0$ is the rate at which cells in G1 transition into S/G2/M, and $k_g > 0$ is the rate at which cells in S/G2/M undergo mitosis to produce two cells in G1 [41]. The solution of Equations (3)–(4), $r(x,t)$ and $g(x,t)$, are taken to represent the (continuous) underlying mean densities, while $r_i^{\mathrm{o}}, g_i^{\mathrm{o}}$, introduced in the previous section, represent noisy observations of these at the corresponding spatiotemporal measurement locations $s_i = (x_i, t_i)$.

This model has four parameters: $\theta = (D_r, D_g, k_r, k_g)$. We will use the notation $r(x,t,\theta)$ and $g(x,t,\theta)$ when we need to be explicit about the dependence of the solutions on the parameters, and denote the combined solution at $s = (x,t)$ for parameters θ by $y(s,\theta)$. Given suitable initial conditions, $r(x,0)$ and $g(x,0)$, and parameter values θ, Equations (3)–(4) can be solved numerically (Supplementary Material).

Despite the abundance of experimental data in Figure 1 and the apparent simplicity of Equations (3)–(4), it is unclear whether these experimental data are sufficient to provide reasonable estimates of the four parameters: $\theta = (D_r, D_g, k_r, k_g)$. Therefore, we also consider a simpler model whereby we set $D_r = D_g = D$, implying that cells in G1 phase diffuse at the same rate as cells in S/G2/M phase [42]. This simpler model is characterised by just three unknown parameters: $\theta = (D, k_r, k_g)$.

4. Identifiability analysis

We consider two approaches for identifiability analysis. The first is based on Bayesian inference and MCMC, and follows the approach outlined by Hines et al. [15]. Similarly, Seikmann et al. [16] discuss MCMC for identifiability analysis. The second approach is based on profile likelihood and follows Raue et al. [17]. Both these approaches have been considered in the context of identifiability analysis of non-spatial models, such as ordinary differential equation or hidden Markov models, and, as we will show, the same ideas apply here for models of spatiotemporal processes.

As discussed by Raue et al. [40], estimation results from MCMC and profile likelihood are typically similar in the case of identifiable parameters (and in the low-dimensional setting), but can be very different in the presence of non-identifiability. In particular, Raue et al. [40] demonstrate that results from MCMC can be potentially misleading in the presence of non-identifiability. On the other hand, Hines et al. [15] and Seikmann et al. [16] argue that appropriate MCMC diagnostics can indicate the presence of non-identifiability and potentially unreliable inferences. Hence we use both methods and consider the extent to which they agree.

4.1. Parameter bounds

Before outlining the identifiability methodologies, we note that the parameters in Equations (3)–(4) have a physical interpretation and we can formulate some biologically motivated bounds for the parameter estimates. Practical identifiability can be evaluated by comparing how precise interval estimates are relative to these simple bounds.

With regards to cell diffusivity, previous studies have estimated the diffusivity of melanoma cells in similar experiments and found that typical values are often less than 1000 μm^2 /h [43] so we take conservative bounds on D_r and D_g to be $0 < D_r, D_g < 2000$ μm^2 /h. Bounds on k_r and k_g can be inferred from previous experimental measurements. The duration of time 1205Lu melanoma cells remain in the G1 phase varies between 8 - 30 h, whereas the time 1205Lu melanoma cells remain in the S/G2/M phase varies between 8 - 17 h [22]. These measurements imply $0.033 < k_r < 0.125$ /h and $0.059 < k_g < 0.125$ /h, so we take conservative bounds on k_r and k_g to be $0 < k_r, k_g < 0.2$ /h.

4.2. Bayesian inference

Following Hines et al. [15] we take a Bayesian approach to assess parameter identifiability. Bayesian inference relies on Bayes' theorem, written here as

$$p(\theta \mid y^{\mathrm{o}}) \propto p(y^{\mathrm{o}} \mid \theta) p(\theta), \tag{5}$$

where $p(y^{\mathrm{o}} \mid \theta)$ is the *likelihood function*, $p(\theta)$ is the *prior* and $p(\theta \mid y^{\mathrm{o}})$ is the *posterior*. The posterior distribution is the inferential target; this summarises the information about the parameter θ in light of the observed data, y^{o}, and the prior information specified by $p(\theta)$. The likelihood function represents the information contributed by the data; to define this we first assume that the observations y_i^{o} are simply noisy versions of the latent model solutions $y(s_i, \theta)$, and that this observation error is independent and identically-distributed (iid). We further assume for convenience that this noise is additive and normally-distributed, with zero mean and common variance σ^2, giving an explicit likelihood function characterised by

$$y_i^{\mathrm{o}} \mid \theta \sim \mathcal{N}(y(s_i, \theta), \sigma^2 I), \tag{6}$$

where I is the 2×2 identity matrix. As discussed above, $y(s_i, \theta) = (r(x_i, t_i, \theta), g(x_i, t_i, \theta))$ is the combined vector of solutions of Equations (3)–(4) given values for the parameters θ and as evaluated at the observation locations.

We explore the parameter space θ by sampling its posterior distribution using a Metropolis-Hastings MCMC algorithm [15]. In brief, the Markov Chain starts at position θ_i, and a potential move to θ^* is accepted with probability α given by

$$\alpha = \min\left[1, \frac{p(\theta^* \mid y^{\mathrm{o}})}{p(\theta_i \mid y^{\mathrm{o}})}\right]. \tag{7}$$

This Markov chain will tend to move toward regions of high posterior probability, but it also provides a mechanism to move away from local minima in the posterior distribution by allowing transitions to regions of lower posterior probability. The Markov chain produced by this algorithm explores the parameter space in proportion to the posterior probability and provides a finite number of independent, identically distributed samples from the posterior distribution. Proposals in the MCMC algorithm are made by sampling a multivariate normal distribution with zero mean and covariance matrix Σ^2. When $\theta = (D_r, D_g, k_r, k_g)$ we specify $\Sigma^2 = \mathrm{diag}(10^2, 10^2, 10^{-6}, 10^{-6})$, whereas when $\theta = (D, k_r, k_g)$ we specify $\Sigma^2 = \mathrm{diag}(10^2, 10^{-6}, 10^{-6})$.

Poor identifiability in a Bayesian setting, and when using MCMC, is typically characterised by poorly converging chains, label-switching, multimodal or overly-broad distributions, and similar phenomena that can be diagnosed either graphically or by computing various diagnostic statistics [15, 16]. We discuss these diagnostic methods in more detail in the Results section.

4.3. Profile likelihood

Here we describe how a profile likelihood identifiability analysis can be undertaken. This is based on the same likelihood $p(y^{\mathrm{o}} \mid \theta)$ as above, but here we will use the normalised likelihood function, denoted by

$$\mathcal{L}(\theta; y^{\mathrm{o}}) = \frac{p(y^{\mathrm{o}} \mid \theta)}{\sup_\theta p(y^{\mathrm{o}} \mid \theta)}. \tag{8}$$

This is considered as a function of θ for fixed data y^{o}. It is also common to take the likelihood as only defined up to a constant, but here we fix the proportionality constant by using the normalised likelihood function as above.

We assume our full parameter θ can be partitioned into an *interest* parameter ψ and *nuisance* parameter λ, i.e. $\theta = (\psi, \lambda)$. Then the profile likelihood for the interest parameter ψ can be written as [28, 39]

$$\mathcal{L}_p(\psi; y^{\mathrm{o}}) = \sup_\lambda \mathcal{L}(\psi, \lambda; y^{\mathrm{o}}). \tag{9}$$

In the above, λ is 'optimised out' for each value of ψ, and this implicitly defines a function $\lambda^*(\psi)$ of optimal λ values for each value of ψ.

It is often convenient computationally to work with log-likelihood functions, and this does not change the induced ordering of parameter values. Given the assumption of normally distributed errors as introduced above, profiling using log-likelihood functions reduces to solving a series of nonlinear least squares problems, one for each value of the parameter of interest. We implement this optimisation using MATLABs *lsqnonlin* function with the 'trust-region-reflective' algorithm to implement bound constraints [44]. For each value of the interest parameter, taken over a sufficiently fine grid, the nuisance parameter is optimised out; the previous optimal value is used as the starting guess for the next optimisation problem for efficiency. Uniformly spaced grids of 50 points, defined between lower and upper bounds based on those given above, are used. Results are plotted in terms of the normalised profile likelihood functions; cubic interpolations are used to define the profiles for all parameter values.

The likelihood function is often characterised as representing the information that the data contains about the parameters, and the relative likelihood for different parameter values as indicating the relative evidence for these parameter values [39, 45, 46]. As such, a flat profile is indicative of non-identifiability, therefore a lack of information in the data about a parameter [17]. In general, the degree of curvature is related to the inferential precision [17, 18, 40]. Likelihood-based confidence intervals can be formed by choosing a threshold-relative profile likelihood value, which can be approximately calibrated via the chi-square distribution (or via simulation). We use a threshold of 0.15 as a reference, which corresponds to an approximate 95% confidence interval for sufficiently regular problems [39]. The points of intersection were determined using the interpolated profile likelihood functions.

5. Results and Discussion

Here we consider the results of the above methods of identifiability analysis under the two scenarios introduced previously. In Scenario 1 we consider Equations (3)–(4) under the assumption that both subpopulations have the same diffusivity, $D_r = D_g = D$, so that $\theta = (D, k_r, k_g)$. In Scenario 2 we consider the same model without this assumption so that $\theta = (D_r, D_g, k_r, k_g)$. We first consider the Bayesian analysis using MCMC, and then the profile likelihood analysis. Throughout this document all units of diffusivities are μm^2/h and all units for rate parameters are /h.

5.1. Bayesian analysis

Here we consider the results from our Bayesian analysis using MCMC. As we are focused on the question of identifiability, here we use uniform priors on all parameters. This is a natural choice when we are interested in identifiability since we want to focus on the information about the parameter values inherent within the data rather than from some imposed prior [15].

In the first scenario where $D_r = D_g = D$, we see rapid convergence of the Markov chain for D, k_r and k_g (Figure 2a-c). Importantly, after the Markov chain moves away from the initial location, θ_0, it remains within the previously-stated conservative bounds. Additional results (not shown) confirm similar results for different choices of θ_0. A plot matrix representation of the univariate and bivariate marginal densities is given (Figure 2d-i) [47]. The univariate marginals (Figure 2d, g, i) are unimodal and approximately symmetric. The univariate posterior modes are $\bar{\theta} = (705, 0.0266, 0.0727)$, confirming that each component lies within the expected range, and the univariate 95% credible intervals are relatively narrow about the posterior mode. A posterior predictive check [47] with 30 randomly sampled parameter choices from the converged region of the Markov Chain provides visual confidence in our MCMC results and the appropriateness of the model (Figure 2o-r) [47]. Additional comparison of the experimental data and solution of the model parameterised with $\bar{\theta}$ also confirms this (Supplementary Material). In summary, assuming that $D = D_r = D_g$, our model parameters appear to be identifiable according to the Bayesian approach.

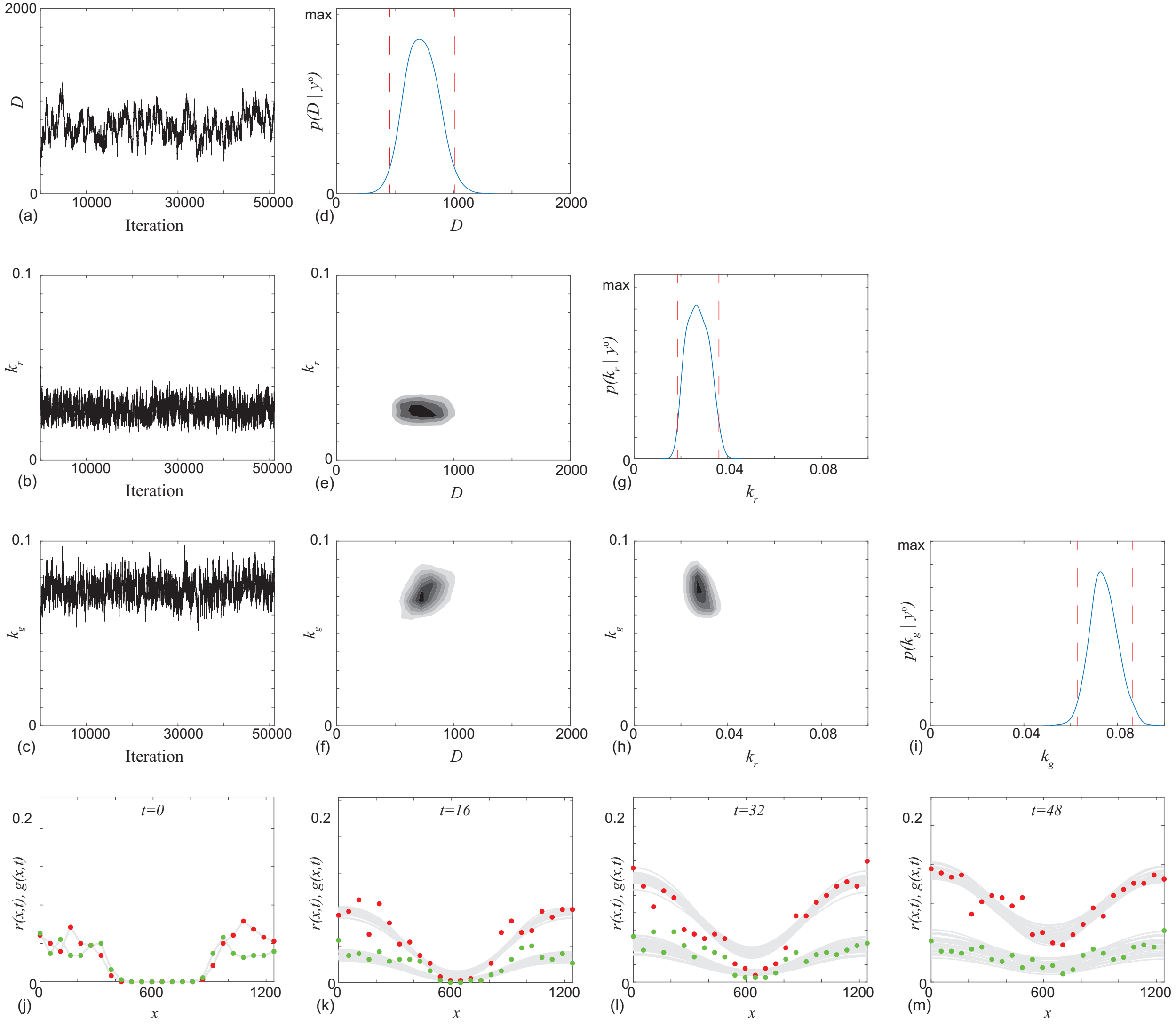

Figure 2: Typical Markov chain iterations, of length 51,000, for D, k_r and k_g in (a)–(c), respectively. In this case the Markov chain is initiated with $\theta_0 = (500, 0.05, 0.05)$, and other choices of initial conditions lead to similar results (not shown). Results in (d)–(i) show a plot matrix representation of the univariate marginals and bivariate marginals estimated using the final 50,000 iterations of the Markov chain in (a)–(c). For the univariate distribution the posterior modes are $\bar{D} = 707$ $\mu\mathrm{m}^2$ /h, $\bar{k}_r = 0.0266$ /h, and $\bar{k}_g = 0.0727$ /h, and the 95% credible intervals are $D \in [455, 1007]$, $k_r \in [0.0187, 0.0362]$ and $k_g \in [0.0628, 0.0727]$. In the univariate marginals the 95% credible intervals are shown in red dashed vertical lines, in the bivariate marginals the region of maximum density is shown in the darkest shade. Results in (j)–(m) superimpose 30 solutions of Equations (3)–(4) where θ is randomly sampled from the Markov Chain. MCMC results use $\sigma = 0.05$.

In the second scenario, where $D_r \neq D_g$, the Markov chains for D_r, D_g, k_r and k_g (Figure 3a-d) are very different to the first scenario. The Markov chains for k_r and k_g remain within the previously-stated conservative bounds. However, the Markov chains for D_r and D_g often move beyond the previously-stated conservative bounds with no obvious convergence. Additional results for different choices of θ_0 (Supplementary Material) indicate that these poor results for D_r and D_g are consistent across a number of choices of θ_0. A plot matrix representation confirms that we obtain well-behaved posterior distributions for k_r and k_g, but that the results for D_r and D_g are problematic since the univariate posterior distributions are not unimodal and the 95% credible intervals are relatively wide. A posterior predictive check with 30 randomly sampled parameter choices from the latter part of the Markov chain are provided for completeness (Figure 2o–r). The posterior predictive check compares model predictions with the experimental data, with additional comparisons using the posterior median (Supplementary Material). However, these comparisons are less useful since they are based on parameters arising from a Markov chain that has not settled to a well-defined posterior. In summary, without assuming $D_r = D_g$, our model parameters are practically non-identifiable according to the Bayesian approach.

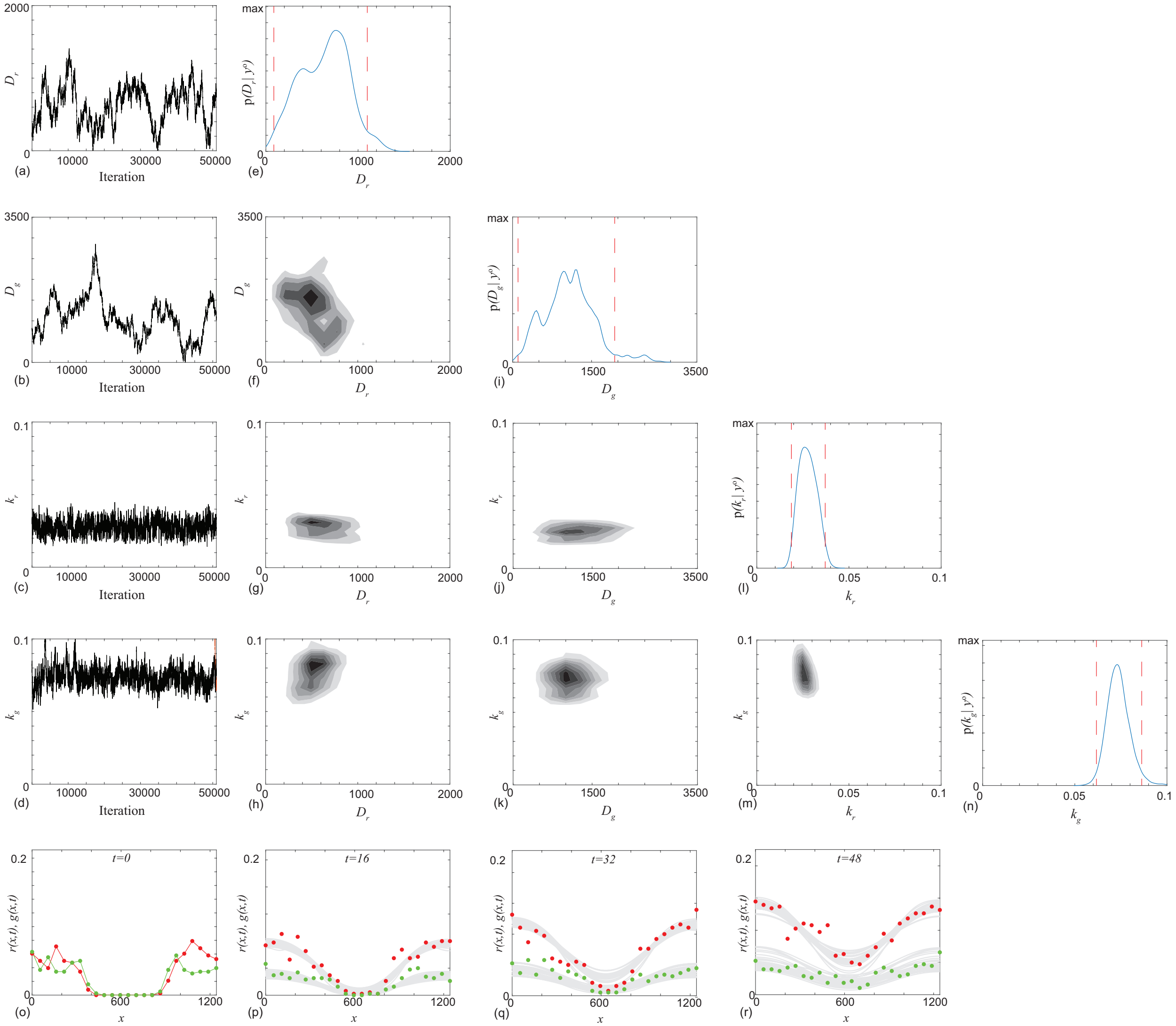

Figure 3: Typical Markov chain iterations, of length 51,000, for D_r, D_g, k_r and k_g in (a)–(d), respectively. In this case the Markov chain is initiated with $\theta_0 = (228, 803, 0.0334, 0.0790)$, and other choices of initial conditions lead to similar results (Supplementary Material). Results in (e)–(n) show a plot matrix representation of the univariate marginals and bivariate marginals estimated using the final 50,000 iterations of the Markov chain in (a)–(d). For the univariate distribution the posterior modes are $\bar{D}_r = 758$ μm^2 /h, $\bar{D}_g = 1203$ μm^2 /h, $\bar{k}_r = 0.0259$ /h, and $\bar{k}_g = 0.0703$ /h. The 95% credible intervals are $D_r \in [86, 1100]$, $D_g \in [104, 1939]$, $k_r \in [0.0188, 0.0371]$ and $k_g \in [0.0617, 0.0863]$. In the univariate marginals the 95% credible intervals are shown in red vertical dashed lines, in the bivariate marginals the region of maximum density is shown in the darkest shade. Results in (o)–(r) superimpose 30 solutions of Equations (3)–(4) where θ is randomly sampled from the Markov Chain. MCMC results use $\sigma = 0.05$.

5.2. *Profile likelihood analysis*

Here we consider the results from our profile likelihood analysis. Each profile corresponds to taking a particular parameter as the interest parameter and the remaining parameters as nuisance parameters to be optimised out. More general partitions into an interest parameter and nuisance parameters are also considered in the Supplementary Material. Here we present results for Scenario 1 (Figure 4a–c) and results for Scenario 2 (Figure 4d–f).

In the first scenario where $D_r = D_g = D$, we see regular shaped profiles with clearly defined peaks for D, k_r and k_g (Figure 4a-c). These results are very similar to those obtained from the Bayesian analysis using MCMC, both in terms of interval estimates and maximum likelihood estimates. The maximum likelihood estimates are $\theta^* = (D^*, k_r^*, k_g^*) = (691, 0.0289, 0.0719)$.

In the second scenario, where $D_r \neq D_g$, we see very similar estimates for k_r and k_g. The diffusivities, D_r and D_g, are much less well-determined, however. In particular, the profile-likelihood-based confidence interval for D_g covers the entire range of the previously-stated conservative bounds, while that for D_r only excludes a negligible part of this region. Thus, while k_r and k_g are well identified, D_r and D_g are practically non-identifiable. These results are again very similar to those obtained from the Bayesian analysis using MCMC, though the diffusivity estimates are even more conservative here. The maximum likelihood estimates are $\theta^* = (D_r^*, D_g^*, k_r^*, k_g^*) = (612, 842, 0.0290, 0.0717)$. We also generated profiles for the difference $D_r - D_g$ (Supplementary Material). The associated interval encompasses a broad range of values, including zero, but also includes very large positive and negative values. This is consistent with a lack of practical identifiability of $D_r - D_g$. Despite the fact that the diffusivities are practically non-identifiable with our relatively abundant experimental data, further experimentation with synthetic data indicates that all parameters are, in principle, identifiable provided sufficient data (Supplementary Material).

In both cases the results of the profile likelihood analysis is consistent with the MCMC-based Bayesian analysis. However, the profiles are far cheaper to compute than full MCMC posterior distributions. For example, each profile here was computed in less than one minute on a standard laptop whereas the MCMC analysis took approximately 30 minutes of computation. Each optimisation for a given value of the target parameter in the profile analysis took less than a second to solve, and profiles can be computed in perfect parallel.

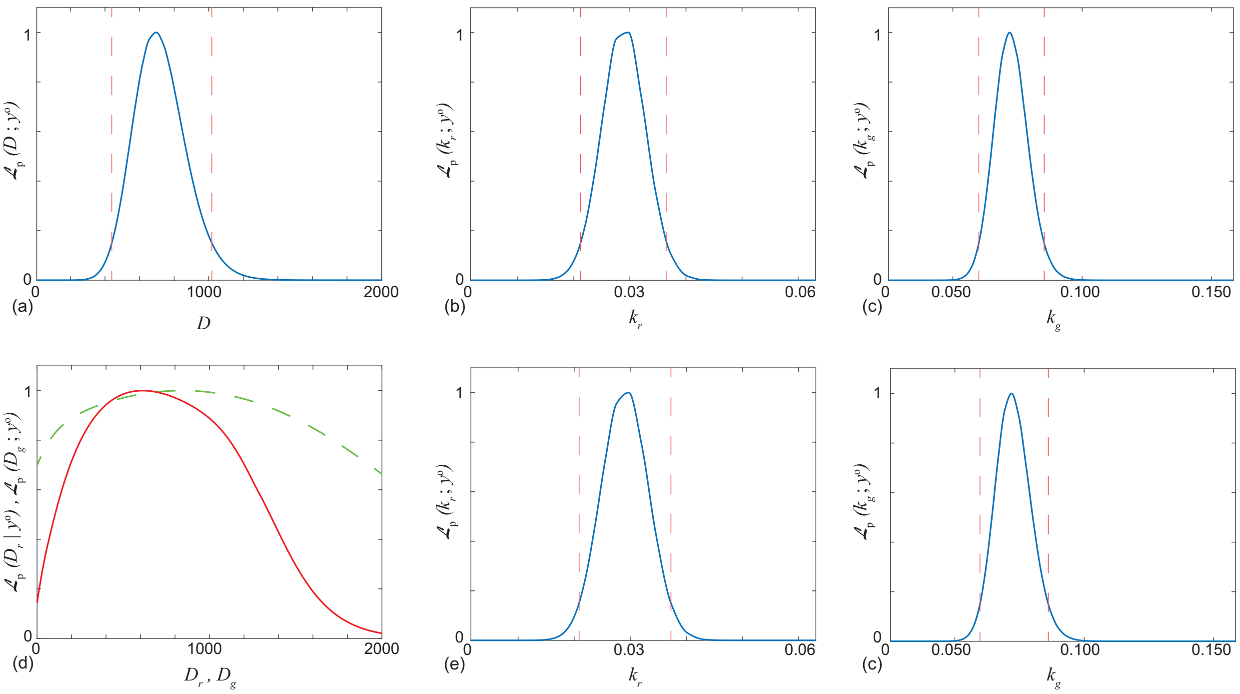

Figure 4: Profile likelihoods. Each profile likelihood is computed using a grid of 50 equally-spaced points; cubic interpolation is used for display and to determine confidence intervals. Approximate 95% confidence intervals are indicated based on a relative likelihood threshold of 0.15 [39]. Results for the first scenario where $D_r = D_g = D$ (a-c), and results for the second scenario where $D_r \neq D_g$ (d-f). In (d) the profile likelihood for D_g and D_r are in red (solid) and green (dashed), respectively. Interval estimates are: $D \in [439, 1018]$, $k_r \in [0.0212, 0.0366]$, $k_g \in [0.0601, 0.0853]$ for $D = D_r = D_g$. Maximum likelihood estimates are: D^*, k_r^*, $k_g^* = 691, 0.0289, 0.0719$ for $D = D_r = D_g$. Interval estimates are: $D_r \in [1.4, 1647]$, $D_g \in [0, 2000]$ (i.e. the whole interval), $k_r \in [0.0209, 0.0373]$, $k_g \in [0.0597, 0.0861]$ for $D_r \neq D_g$. Maximum likelihood estimates are: D_r^*, D_g^*, k_r^*, $k_g^* = 612, 842, 0.0290, 0.0717$ for $D_r \neq D_g$.

6. Conclusions and Outlook

In this work we assess the practical identifiability of a spatiotemporal reaction-diffusion model for a scratch assay where we have access to a relatively large amount of quantitative experimental data. Most scratch assays in the experimental cell biology literature are reported by providing single images at the beginning and the conclusion of the experiment without making measurements of spatial distributions of cell density at multiple time points. In contrast, here we consider detailed cell density measurements of two subpopulations at four time points with a relatively high spatial resolution of cell density at 24 spatial locations. This means that we work with $24 \times 4 \times 2 = 192$ measurements of cell density. Given this comprehensive data set, a key decision when using a mathematical model to describe the experiment is to determine how complicated the model ought to be. We first explored this question through a Bayesian, MCMC-based framework, and then with a profile likelihood-based approach.

MCMC-based Bayesian analysis is by now relatively common in the systems biology literature for temporal processes and temporal models [33]. Bayesian MCMC approaches for spatiotemporal models are less common, but increasingly of interest in the mathematical biology community. While it is established that practical parameter identifiability can often be diagnosed using an MCMC framework, many applications of Bayesian MCMC in the mathematical biology literature never explicitly consider the question of parameter identifiability [34]. Furthermore, MCMC can be misleading in the presence of true non-identifiability [40], as can other methods such as the Bootstrap or Fisher-information-based approaches [18]. Even assuming that MCMC can reliably determine non-identifiability, it can be computationally expensive, especially for models with a high-dimensional parameter space. Thus we also assess identifiability using an optimisation-based, profile likelihood approach. In contrast to past work focusing on temporal data and temporal processes, here we apply this approach to spatiotemporal reaction–diffusion models and spatiotemporal data. Algorithms to implement both the profile likelihood and Bayesian MCMC approaches are available on GitHub.

A key feature of our study is that we consider two slightly different modelling scenarios. In the first scenario we assume $D = D_r = D_g$ so that the diffusivity of cells in G1 phase is identical to the diffusivity of cells in the S/G2/M phase. In this case there are three unknown parameters, $\theta = (D, k_r, k_g)$. In contrast, for the second scenario we do not invoke this assumption, and there are four unknown parameters, $\theta = (D_r, D_g, k_r, k_g)$. Despite the relative abundance of experimental data and the apparent simplicity of the reaction-diffusion model, we find that the parameters are identifiable in the first scenario but are practically non-identifiable in the second scenario.

Our results show that the profile likelihood provides similar results to the Bayesian MCMC approach, with the advantage of being an order of magnitude faster to implement for the problems that we consider. Typical MCMC results in Figures 2-3 require approximately 30 minutes of computation time whereas the profile likelihood results in Figure 4 require approximate one minute to compute on a standard laptop (DELL, Intel Core i7 Processor, 2.8GHz, 16GB RAM). The similarity between the Bayesian MCMC results and the profile likelihood analysis is somewhat expected as here we used flat priors in the Bayesian analysis to assess identifiability, though profiling is based on maximising out nuisance parameters while in the Bayesian approach they are integrated out. Both methods indicate practical non-identifiability of the diffusion coefficients in the more general second scenario. As mentioned, however, profile likelihood has previously been shown to be more reliable in the presence of true non-identifiability [18, 40]; this is an important consideration as the reaction-diffusion models we used here are already relatively simple and neglect certain mechanisms. Potential mechanisms that might be incorporated in an extended model include cell-cell adhesion [8, 48] or directed motion such as chemotaxis [49–51].

Visual inspection of the posterior predictive check in Figure 2 shows that while the overall comparison between the model and the experimental data is reasonable, there are some minor discrepancies, such as in the G1 (red) subpopulation near the centre of the scratch at $t = 32$ h, suggesting some model inadequacy. This observation is consistent with the fact that we neglect some mechanisms that could improve the quality of the match. Had we taken a more standard approach without considering an identifiability analysis, we might have been tempted to extend Equations (3)–(4) to improve the quality of match between the data and the solution of the model. Such extensions, while conceptually straightforward to implement, involve increasing the dimension of the parameter space. We caution against naively implementing such extensions since our identifiability analysis shows that even the simpler models we consider are not always practically identifiable even with the extensive data set that we work with here. A profile likelihood analysis of particular, lower-dimensional *interest parameters* within a more complex model may still be feasible, however.

There are many ways in which our analysis could be extended. A key assumption in our work is that the noise model is Gaussian, and this is a standard assumption in many applications in the mathematical biology literature [10]. However, since we are dealing with count data, a different noise model, such as a multinomial model, would technically be more appropriate. Additional results (not shown) confirm that results obtained using a multinominal noise model do not lead to qualitatively different conclusions than those reported here with the Gaussian noise model. A further extension would be to apply the reaction-diffusion models described in this work to a different, more realistic geometry. For example, experimental data reported by Jin et al. [52] uses a different experimental approach to produce square and triangular shaped scratches. While very similar reaction-diffusion models to those used here could be used to mimic these experiments with different shaped scratches, such models would need to be solved in two spatial dimensions because of the lack of symmetry in the initial condition [53]. This extension would not necessarily increase the dimension of the parameter space, but it would increase the computational effort required to solve the reaction-diffusion model. This increase in computational effort to solve the reaction-diffusion models in a different coordinate system would further amplify the computational advantage of the profile likelihood method over the Bayesian MCMC approach.

Overall, we recommend profile likelihood as a useful screening tool that has great potential in the field of mathematical biology since it can provide rapid insight into practical identifiability. This is consistent with the recommendations of Raue et al. [40] in the context of temporal models and temporal data. As the models and data considered in mathematical biology become increasingly complicated and spatiotemporal in nature, we expect tools for rapid identifiability analysis such as profile likelihood to become correspondingly more important.

Acknowledgements MJS is supported by the Australian Research Council (DP170100474) and the Erskine Fellowship at the University of Canterbury. OJM is supported through the University of Auckland, Faculty of Engineering James and Hazel D. Lord Emerging Faculty Fellowship. REB acknowledges the Royal Society Wolfson Research Merit Award, the Leverhulme Trust Research Fellowship and the BBSRC (BB/R000816/1).

[1] Simpson MJ, Treloar KK, Binder BJ, Haridas P, Manton KJ, Leavesley DI, McElwain DLS, Baker RE. 2013. Quantifying the roles of cell motility and cell proliferation in a circular barrier assay. Journal of the Royal Society Interface. 10. 20130007.

[2] Maini PK, McElwain DLS, Leavesley DI. 2004. Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. Tissue Engineering. 10. 475-482.

[3] Liang CC, Park AY, Guan J-L. 2007. *In vitro* scratch assay: a convenient and inexpensive method for analysis of cell migration *in vitro*. Nature Protocols. 2. 329-333.

[4] Grada A, Otero-Vinas M, Prieto-Castrillo F, Obagi Z, Falanga V. 2016. Research techniques made simple: analysis of collective cell migration using the wound healing assay. Journal of Investigative Dermatology. 137. e11-e16.

[5] Murray JD. 2002. Mathematical biology I: An introduction. Heidelberg: Springer.

[6] Sherratt JA, Murray JD. 1990. Models of epidermal wound healing. Proceedings of the Royal Society of London Series B. 241. 29-36.

[7] Sengers BG, Please CP, Oreffo ROC. 2007. Experimental characterization and computational modelling of two-dimensional cell spreading for skeletal regeneration. Journal of the Royal Society Interface. 4. 1107-1117.

[8] Nardini JT, Chapnick DA, Liu X, Bortz DM. 2016. Modeling keratinocyte wound healing dynamics: Cell-cell adhesion promotes sustained collective migration. Journal of Theoretical Biology. 400. 103-117.

[9] Bobadilla AVP, Arévalo J, Sarró E, Byrne HM, Maini PK, Carraro T, Balocco S, Meseguer A, Alarcón T. 2019. *In vitro* cell migration quantification method for scratch assays. Journal of the Royal Society Interface. 16. 20180709.

[10] Warne DJ, Baker RE, Simpson MJ. 2019. Using experimental data and information criteria to guide model selection for reaction-diffusion problems in mathematical biology. Bulletin of Mathematical Biology. 81. 1760-1804.

[11] Johnston ST, Ross RV, Binder BJ, McElwain DLS, Haridas P, Simpson MJ. (2016). Quantifying the effect of experimental design choices for in vitro scratch assays. Journal of Theoretical Biology. 400. 19-31.

[12] Audoly S, Bellu G, D'Angiò L, Saccomani MP, Cobelli C. 2001. Global identifiability of nonlinear models of biological systems. IEEE Transactions on Biomedical Engineering. 48. 55-65.

[13] Chis O-T, Banga JR, Balsa-Canto E. 2011. Structural identifiability of systems biology models: a critical comparison of methods. PLoS ONE. 6. e27755.

[14] Cobelli C, Distefano JJ. 1980. Parameter and structural identifiability concepts and ambiguities: a critical review and anaysis. American Journal of Physiology. Regulatory and Integrative and Comparitive Physiology. 8. R7-R24.

[15] Hines KE, Middendorf TR, Aldrich RW. 2014. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. Journal of General Physiology. 143. 401.

[16] Siekmann I, Sneyd J, Crampin EJ. 2012. MCMC can detect nonidentifiable models. Biophysical Journal. 103. 2275-2286.

[17] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 25(15). 1923-1929.

[18] Fröhlich F, Theis FJ, Hasenauer J. 2014. Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more. International Conference on Computational Methods in Systems Biology. 61-72. Springer.

[19] Villaverde AF, Tsiantis N, Banga JR. 2019. Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models. Journal of the Royal Society Interface. 16(156). 20190043.

[20] Campbell DA, Chkrebtii O. 2013. Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimate. Mathematical Biosciences. 246(2). 283-292.

[21] Sakaue-Sawano A, Kurokawa H, Morimura T, Hanyu A, Hama H, Osawa H, Kashiwagi S, Fukami K, Miyata T, Miyoshi H, Imamura T, Ogawa M, Masai H, Miyawaki A. 2008. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. Cell. 132. 487-498.

[22] Haass NK, Beaumont KA, Hill DS, Anfosso A, Mrass P, Munoz MA, Kinjyo I, Weninger W. 2014. Realtime cell cycle imaging during melanoma growth, invasion, and drug response. Pigment Cell & Melanoma Invasion. 27: 764-776.

[23] Bickel PJ, Doksum KA. 2015. Mathematical statistics: Basic ideas and selected topics. Vol 1. CRC Press.

[24] Greenland S. 2011. Identifiability. International encyclopedia of statistical science. 645-645.

[25] Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions of Automatic Control. 19(6). 716-723.

[26] Shmueli G. 2010. To explain or to predict? Statistical Science. 25(3). 289-310.

[27] Maclaren OJ, Nicholson R. 2019. What can be estimated? Identifiabiliy, estimability, casual infernce and ill-posed inverse problems. arXiv. https://arxiv.org/abs/1904.02826.

[28] Cox DR. 2006. Principles of statistical inference. Cambridge University Press.

[29] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. 2016. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. 31(4). 337-350.

[30] Rothman KJ, Greenland S. 2018. Planning study size based on precision rather than power. Epidemiology. 29(5). 599-603.

[31] Bland JM. 2009. The Tyranny of power: Is there a better way to calculate sample size? BMJ. 339. b3985.

[32] Warne DJ, Baker RE, Simpson MJ. 2019. Simulation and inference algorithms for biochemical stochastic reaction networks: from basic concepts to state-of-the-art. Journal of the Royal Society Interface. 16, 20180943.

[33] Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface. 6. 187-202.

[34] Vo BN, Drovandi CC, Pettitt AN, Simpson MJ. 2015. Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. Mathematical Biosciences. 263. 133-142.

[35] Browning AP, Haridas P, Simpson MJ. 2019. A Bayesian sequential learning framework to parameterise continuum models of melanoma invasion into human skin. Bulletin of Mathematical Biology. 81. 676-698.

[36] Maclaren OJ, Parker A, Pin C, Carding SR, Watson A, Fletcher AG, Byrne HM, Maini PK. 2017. A hierarchical Bayesian model for understanding the spatiotemporal dynamics of the intestinal epithelium. PLoS Computational Biology. 13(7). e1005688.

[37] Lambert B, MacLean AL, Fletcher AG, Combes AN, Little MH, Byrne HM. 2018. Bayesian inference of agent-based models: a tool for studying kidney branching morphogenesis. Journal of Mathematical Biology. 76. 1673-1697.

[38] Wilkinson RD. 2013. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Statistical Applications in Genetics and Molecular Biology. 12. 129-141

[39] Pawitan Y. 2001. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press.

[40] Raue A, Kreutz C, Theis FJ, Timmer J. 2013. Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 371. 20110544.

[41] Vittadello ST, McCue SW, Gunasingh G, Haass NK, Simpson MJ. 2018. Mathematical models for cell migration with real-time cell cycle dynamics. Biophysical Journal. 114. 1241-1253.

[42] Vittadello ST, McCue SW, Gunasingh G, Haass NK, Simpson MJ. 2019. Examining go-or-grow using fluorescent cell-cycle indicators and cell cycle-inhibiting drugs. bioRxiv. doi https://doi.org/10.1101/797142.

[43] Treloar KK, Simpson MJ, Haridas P, Manton KJ, Leavesley DI, McElwain DLS, Baker RE. 2013. Multiple types of data are required to identify the mechanisms influencing the spatial expansion of melanoma cell colonies. BMC Systems Biology. 7. 137.

[44] Mathworks. 2019. lsqnonlin. Solve nonlinear least-squares. https://au.mathworks.com/help/optim/ug/lsqnonlin.html. (Accessed November 2019).

[45] Edwards AW. 1992. Likelihood. Baltimore: Johns Hopkins University Press. Expanded ed.

[46] Royall R. 2017. Statistical evidence: a likelihood paradigm. Routledge.

[47] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. Bayesian data analysis, 3rd edition. Chapman & Hall, London.

[48] Armstrong NJ, Painter KJ, Sherratt JA. 2006. A continuum approach to modelling cell-cell adhesion. Journal of Theoretical Biology. 243. 98-113.

[49] Keller EF, Segal LA. 1971. Model for chemotaxis. Journal of Theoretical Biology. 30. 225-234.

[50] Hillen T, Painter KJ. 2009. A user's guide to PDE models for chemotaxis. Journal of Mathematical Biology. 58. 183-217.

[51] Simpson MJ, Landman KA, Newgreen DF. 2006. Chemotactic and diffusive migration on a nonuniformly growing domain: numerical algorithm development and applications. Journal of Computational and Applied Mathematics. 192. 282-300.

[52] Jin W, Lo K-Y, Chou S-E, McCue SW, Simpson MJ. 2018. The role of initial geometry in experimental models of wound closing. Chemical Engineering Science. 179. 221-226.

[53] Simpson MJ, Landman KA, Hughes BD. 2009. Multi-species simple exclusion processes. Physica A. 388. 399-406.