Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 16(6): 7299–7326.
DOI: 10.3934/mbe.2019365
Received: 28 March 2019
Accepted: 01 August 2019
Published: 09 August 2019

Research article

Estimation of probability distributions of parameters using aggregate population data: analysis of a CAR T-cell cancer model

Celia Schacht[1], Annabel Meade[1], H.T. Banks[1,*], Heiko Enderling[2] and Daniel Abate-Daga[2]

[1] Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695, USA

[2] H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA

* **Correspondence:** Email: htbanks@ncsu.edu; Tel: +19195158968; Fax: +19195151636.

Abstract: In this effort we explain fundamental formulations for aggregate data inverse problems requiring estimation of probability distribution parameters. We use as a motivating example a class of CAR T-call cancer models in mice. After ascertaining results on model stability and sensitivity with respect to parameters, we carry out first elementary computations on the question how much data is needed for successful estimation of probability distributions.

Keywords: aggregate data; CAR T-cell therapy; cancer model; inverse problems; design of experiments

1. Introduction

In the mathematical modeling of physical and biological systems, situations often arise in which some facet of the underlying dynamics (in the form of a parameter) is not constant but rather may be distributed probabilistically within the system or across the population under study. While traditional inverse problems involve the estimation, given a set of data/observations, of a fixed set of parameters contained within some finite dimensional admissible set, models with parameters distributed across the population require the estimation of a probability measure or distribution over the set of admissible 'parameters'. The techniques for such measure estimation (along with their theoretical justifications) are widely scattered throughout the literature in applied mathematics and statistics, often with few cross references to related ideas; reviews are given in [1, 2]. Of course, it is highly likely that individual parameters might vary from one individual to the next within the sampled population. Thus, our goal in this case is to use the sample of individuals to estimate the probability measure describing the distribution of certain parameters in the full population. In some situations (e.g., in certain pharmacokinetics and/or pharmacodynamics examples), one is able to follow each individual separately and

collect longitudinal time course data for each individual. In other investigations, situations arise in which one is only able to collect *aggregate data*. Such might be the case, for example, in marine or insect catch and release experiments in which one cannot be certain of measuring the same individual multiple times. In other situations, one might be required to sacrifice each individual in the process of collecting the data (precisely the situation in the data we present below). In these cases, one does not have individual longitudinal data, but rather histograms showing the aggregate number of individuals sampled from the population at a given time, having a given size/weight or other characteristic of interest ([3] and Chapter 9 of [4]). The goal then is to estimate the probability distributions describing the variability of the parameters across the population.

In our problem below, while one again has a proposed individual model, the data collected *cannot* be identified with individuals and is considered to be *sampled longitudinally* from the *aggregate population*. It is worth noting that special care must be taken in this case to identify the model such as that introduced below as an individual model in the sense that it describes an individual subpopulation. That is, we have all 'individuals' (i.e., shrimp [5] or mosquitofish [3, 6, 7] or cancerous mice [8, 9]) described by the model sharing a common growth/birth/death rate function. Mathematically, here we define the 'individual' in terms of the underlying parameters, using 'individual' to describe the unit characterized by a single parameter set (tumor expansion rate, excretion rate, growth/birth/death rate, damping rate, relaxation times, etc.).

One can also distinguish two generic estimation problems. In the example presented in this manuscript, we consider the case, as in a structured density model, that one has a mathematical model for individual dynamics but only aggregate data (we refer to this as the *individual dynamics/aggregate data* problem). The second possibility–that one has only an aggregate model (i.e., the dynamics depend explicitly on a distribution of parameters across the population) with aggregate data is not examined in this manuscript (we refer to this as the *aggregate dynamics/aggregate data* problem). Such examples arise in electromagnetic models with a distribution of polarization relaxation times for molecules (e.g., [10, 11]); in biology with HIV cellular models [12–14]; and in wave propagation in viscoelastic materials such as biotissue [15–18]. The measure estimation problem for such examples is sufficiently similar to the individual dynamics/aggregate data situation and accordingly we do not discuss aggregate dynamics models here.

In the generic estimation problems mentioned above, the underlying goal is the determination of the probability measure which describes the distribution of parameters across all members of the population. Thus, two main issues are of interest. First, a sensible framework (the *Prohorov Metric Framework* as we have developed it–see Chapter 14 of [19] and Chapter 5 of [2]) must be established for each situation so that the estimation problem is mathematically meaningful. Thus we must decide what type of information and/or estimates are desired (e.g., mean, variance, complete distribution function, etc. for the tumor size) and determine how these decisions will depend on the type of data available. Second, we must examine what mathematical techniques are available for the computation of such estimates. Because the space of probability measures is an infinite dimensional space (again a mathematical notion), we must make some type of finite dimensional approximations so that the estimation problem is amenable to convergent computations. A thorough discussion of associated mathematical, statistical and computational issues can be found in [2].

We now turn to two important questions: the first is how to deal with inverse problems (parameter estimation) for dynamical systems when longitudinal data does not exist? A second question that we

shall consider below entails how to efficiently design experiments to collect data (i.e., how much data and when to collect it) that is necessary to validate models in such aggregate data situations. Our efforts here are motivated by a set of data collected with NSG (NOD/SCID/GAMMA), or NOD.Cg-PrkdcscidIl2rgtm1Wjl/SzJ mice injected with with cancer and then sequentially sacrificed to collect the needed data about T-cell counts. The resulting aggregate data is absolutely common in biological experiments where data collection requires sacrifice of the subjects. Colleagues at Moffitt Cancer Center have carried out preliminary trial experiments (data collection already completed) as follows: At $t = -14$ (14 days before the trial begins), NSG mice are injected with cancer, which should take 12–14 days to begin growing. At $t = 0$ (as determined by tumor volume) the mice are further injected with chimeric antigen receptor [8] or CAR T-cells (engineered T-cells to specifically target cancer cells). Mice are divided into four groups with different treatments. Autopsies are performed on each of 5 mice sacrificed at $t = 5$, 10, and 15 days, respectively, to determine the concentration of the engineered T-cells within the blood, spleen, and within the tumor. Because of nature of data collection, longitudinal data is not possible as mice must be killed for data to be collected.

In the present investigation we begin by describing an *individual* model (as opposed to an aggregate model) to describe T-cell counts in an individual cancerous mouse. We first perform stability analysis on the model to understand better its behavior, and carry out sensitivity analysis to identify the most 'important' parameters. Next, we describe the *aggregate* model associated with the individual model, and some techniques to estimate the probability distribution over a chosen parameter. Finally, we simulate the Moffitt Cancer Center data set and demonstrate that parameters cannot be properly estimated with only a handful of time points.

2. Individual model

2.1. Model description

We use the system of ordinary differential equations,

$$\frac{dT}{dt} = \rho\beta_{Ext}B - \beta_{TB}T \tag{2.1}$$

$$\frac{dB}{dt} = (\beta_{TB}T - \rho\beta_{Ext}B) + (\beta_{SB}S - \beta_{Ext}B) \tag{2.2}$$

$$\frac{dS}{dt} = \beta_{Ext}B - \beta_{SB}S, \tag{2.3}$$

based on simple mass balance as described in [4], [20], and [21] and depicted in Figure 1 to model the flow of T-cells in the tumor, T, T-cells in the blood, B, and T-cells in the spleen, S, in a cancerous body. The number of T-cells in the blood, B, travel to the tumor, T, at a rate $\rho_{Tx} = \rho\beta_{Ext}$. For each T-cell that leaves the blood, there may be a transient expansion, ρ, in the tumor due to antigen recognition. If there is no antigen recognition, $\rho_{Tx} = \beta_{Ext}$ and $\rho = 1$, and if there is any antigen recognition, $\rho_{Tx} > \beta_{Ext}$ and $\rho > 1$. The T-cells in the tumor, T, then flow back to the blood, B, at a rate β_{TB}. T-cells in the blood flow to the spleen, S, at a rate $\beta_{BS} = \beta_{Ext}$. That is, the flow rate of T-cells out of the blood to the spleen is the same as the exit rate of T-cells from the blood to the tumor if $\rho = 1$ and there is no antigen recognition. Then the T-cells flow from the spleen back to the blood at a rate β_{SB}. Our four parameters of interest are thus ρ, β_{Ext}, β_{TB}, and β_{SB}, which are all strictly positive. In order to solve the system

of ODE's the initial number of T-cells in the tumor, $T_0 = T(t_1)$, the blood $B_0 = B(t_1)$, and the spleen $S_0 = S(t_1)$ also need to be specified.

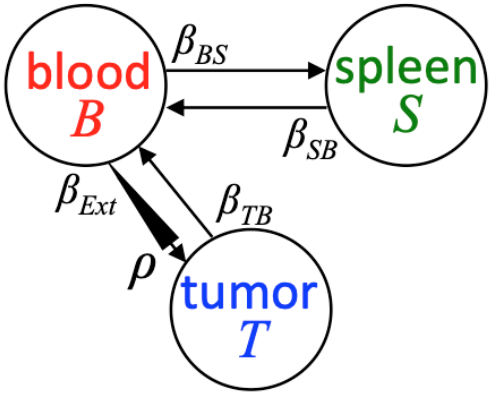


Figure 1. Schematic of the simple compartmental model described in (2.1), (2.2), and (2.3).

In the experiment of interest, cancerous mice are separated into four treatment categories: untransduced T-cells (UT), chimeric antigen receptor therapy (CAR), CAR treatment with added CXCR1 chemokine receptors (CAR+CXCR1), and CAR treatment with added CXCR2 chemokine receptors (CAR+CXCR2). Each treatment has a different effect on antigen recognition in the tumor, or parameter ρ. The parameters and their values are listed in Table 1 and are partially motivated by reference to other tumor-related experiments [9]. The ρ values for each treatment are estimations around which we can study the behavior for each system. The T-cell movement rates, β_{Ext}, β_{TB}, and β_{SB}, are estimated based on knowledge that the exit spleen rate is considerably lower than the exit tumor rate, such that $\beta_{Ext} > \beta_{TB} > \beta_{SB}$. It is also known that the initial T-cell counts in the tumor and spleen, T_0 and S_0, respectively, are zero, and the initial T-cell count in the blood, B_0, ranges from 0 to 10 million. Thus, $B_0 = 10^6$ is chosen, and we set $T_0 = B_0 = 0$

Table 1. Parameters, initial conditions, and their descriptions in equations (2.1)–(2.3).

$\boldsymbol{\theta}$	**Definition**	**Chosen Values**	**Units**
β_{Ext}	Rate at which T-cells exit blood	0.01	1/day
β_{TB}	Rate at which T-cells exit tumor and enter blood	0.001	1/day
β_{SB}	Rate at which T-cells exit spleen and enter blood	0.0001	1/day
$\boldsymbol{\rho}$	**Transient expansion factor of T-cells in the tumor**	**–**	**1**
ρ_{UT}	With no treatment	14	1
ρ_{CAR}	With CAR treatment	15	1
ρ_{CXCR1}	With CAR+CXCR1 treatment	20	1
ρ_{CXCR2}	With CAR+CXCR2 treatment	30	1
T_0	Initial condition of T at the first time point ($T_0 = T(t_1)$)	0	T-cells
B_0	Initial condition of B at the first time point ($B_0 = B(t_1)$)	10^6	T-cells
S_0	Initial condition of S at the first time point ($S_0 = S(t_1)$)	0	T-cells

2.2. Stability analysis

Before comparing data to a mathematical model, it is important to understand the behavior of the mathematical model, especially the limiting behavior. Since the values in Table 1 are set arbitrarily for simulation purposes, ideally we want to understand the behavior of the mathematical model for any realistic parameter values (i.e., values of β_{Ext}, β_{TB}, etc.). Although in reality we will not observe the number of T-cells for longer than a few days, it is important to understand what happens to these state values in the limit as time becomes large. Equations (2.1), (2.2), and (2.3) make up the first-order linear homogeneous system of differential equations

$$\frac{d\boldsymbol{X}}{dt} = J\boldsymbol{X}(t) \tag{2.4}$$

where $\boldsymbol{X}(t) = [T(t), B(t), S(t)]^T$ is a 3×1 vector and

$$J = \begin{bmatrix} -\beta_{TB} & \rho\beta_{Ext} & 0 \\ \beta_{TB} & -\rho\beta_{Ext} - \beta_{Ext} & \beta_{SB} \\ 0 & \beta_{Ext} & -\beta_{SB} \end{bmatrix}. \tag{2.5}$$

is a 3×3 matrix sometimes referred to as the Jacobian matrix [22]. Theoretically, the solution of this system can be found from the eigenvalues and eigenvectors of J (see Chapter 3 of [23]), but for arbitrary parameter values (β_{TB}, β_{SB}, etc.), these may be difficult or impossible to find. However, by examining the eigenvalues, we learn about the stability of the system.

The characteristic equation of J is

$$\lambda^3 + m\lambda^2 + q\lambda = 0 \tag{2.6}$$

where

$$\begin{aligned} m &= \beta_{TB} + \beta_{SB} + \beta_{Ext}(\rho + 1) \\ q &= \beta_{TB}\beta_{SB} + \beta_{TB}\beta_{Ext} + \beta_{SB}\beta_{Ext}\rho. \end{aligned}$$

Thus, one of the eigenvalues is $\lambda_1 = 0$, and by Decartes' Rule of Signs [24], the other two eigenvalues λ_2 and λ_3 are either complex or negative (all parameter values defined in Table 1 are strictly positive). However, the discriminant

$$m^2 - 4q = (\beta_{TB} - \beta_{SB} + \beta_{Ext}(\rho - 1))^2 + 4\beta_{Ext}^2\rho$$

is strictly positive, so λ_2 and λ_3 are both real and negative. Hence, the solution to the system takes the form

$$X(t) = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 e^{\lambda_2 t} + c_3\boldsymbol{v}_3 e^{\lambda_3 t}$$

where c_1, c_2, and c_2 are constants of integration determined by the initial conditions T_0, B_0, and S_0, and $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$ are the eigenvectors corresponding to $\lambda_1 = 0$, λ_2, and λ_3. Hence, as $t \longrightarrow \infty$, the solution of the mathematical model approaches the positive steady state $\boldsymbol{X}(t) \longrightarrow c_1\boldsymbol{v}_1$. The eigenvector $\boldsymbol{v}_1 = [\beta_{SB}\beta_{Ext}\rho, \beta_{TB}\beta_{SB}, \beta_{TB}\beta_{Ext}]^T$ is easily found, but λ_2, λ_3, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$, and the constants c_1, c_2, and c_3 are complicated algebraically and are not necessary to examine steady state behavior.

Under the biologically realistic conditions that $\beta_{Ext} > \beta_{TB} > \beta_{SB} > 0$ and $\rho \geq 1$, the model has two different cases of long term behavior, found through $\boldsymbol{v}_1$:

Tumor-Ignoring Case: If $\rho\beta_{SB} \leq \beta_{TB}$, then $\bar{B} < \bar{T} \leq \bar{S}$.
Tumor-Targeting Case: If $\rho\beta_{SB} > \beta_{TB}$, then $\bar{B} < \bar{S} < \bar{T}$,

where $\lim_{t\to\infty} \boldsymbol{X}(t) = \bar{\boldsymbol{X}} = [\bar{T}, \bar{B}, \bar{S}]^T$ is the steady state of the mathematical model. In the Tumor-Targeting Case, more T-cells are going to the site of the tumor, which is ideal for the patient, but in the Tumor-Ignoring Case, T-cells are either favoring the spleen or treating the spleen and tumor equally. At the current parameter values in Table 1, the Tumor-Ignoring Case occurs when $\rho \leq 10$, and the Tumor-Targeting Case occurs when $\rho > 10$. Thus, at all hypothesized values of the transient expansion factor ρ, T-cells should target the tumor in the long term. Using values from Table 1 with $\rho = 2$ for the Tumor-Ignoring Case and $\rho = 15$ for the Tumor-Targeting Case, Figure 2 below captures this behavior.

Figures 2a and 2b display short and long-term solution behavior for the solutions of the system in the Tumor-Ignoring case, when $\rho = 2$. In this situation, the expansion parameter ρ is less than ideal, which means that the T-cells are not being shuttled to the tumor in a significant way. Figure 2b shows that in the long-term, T-cells in the blood will decrease to zero, and although they increase to an extent in the tumor, they eventually lose numbers and go to the spleen. Figure 2a shows this short-term behavior, where we can see the relatively slow decay of T-cells in the blood and relatively slow growth of T-cells in the tumor and spleen.

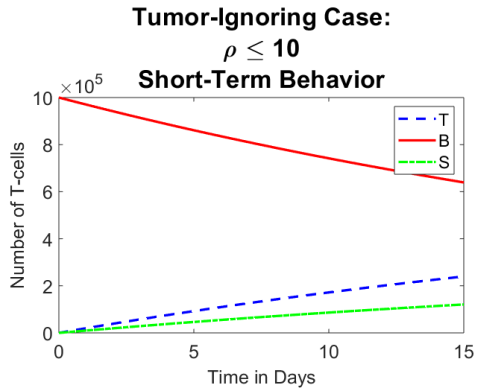
If we look at short and long-term cases in which $\rho > 10$, the Tumor-Targeting case, we see a notable change in behavior of the system. In Figure 2c, we see that T-cells leave the blood and are shuttled to the tumor at a much higher rate than in the Tumor-Ignoring case, while T-cells in the spleen grow very slowly. In the long-term, as seen in Figure 2d, the behavior is indeed very dramatic initially, and then levels out. T-cells in the blood, again, go to zero, while T-cells in the tumor increase significantly and level out, with T-cells in the spleen remaining low over time. This behavior corresponds with a value of ρ that reinforces the Tumor-Targeting behavior: when $\rho > 10$, the body sends the T-cells to the unwanted tumor, which is what we would expect in a cancer treatment. Note that since T-cell movement in the body occurs relatively slowly, T, B, and S take a long time to reach their steady states at biologically relevant parameter values. Thus, using parameter values listed in Table 1, T, B, and S do not reach their steady state until $t > 10,000$ days.

2.3. Parameter selection via local sensitivity analysis

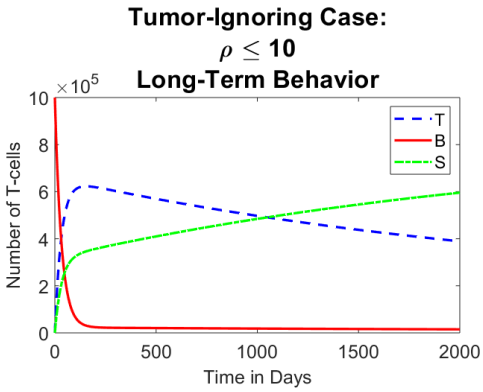
Statistical significance in an inverse problem (fitting data to a mathematical model) depends largely on the sensitivity of the parameter chosen to be estimated. Utilizing sensitivity analysis, we determine which parameters are most significant in affecting the behavior of the model. That is, parameters with high sensitivities dramatically affect the solution, as the observations (T-cell concentrations in the blood, B, tumor, T, and spleen, S) are most sensitive to those parameters, while parameters with low sensitivities have little influence on the model. The sensitivity of observation f to parameter estimates θ is

$$\chi(\theta, t) = \frac{\partial f(t;\theta)}{\partial \theta} \tag{2.7}$$

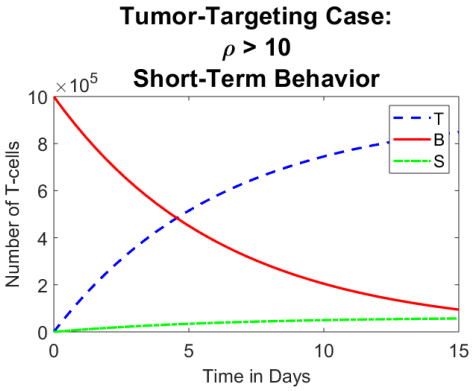
where t is time, and $f = T$, $f = B$, or $f = S$, and θ is one of the parameters defined in Table 1 ($\theta = \beta_{Ext}$ or $\theta = \beta_{TB}$ or etc.). Since many of the parameters have different orders of magnitude (for example,



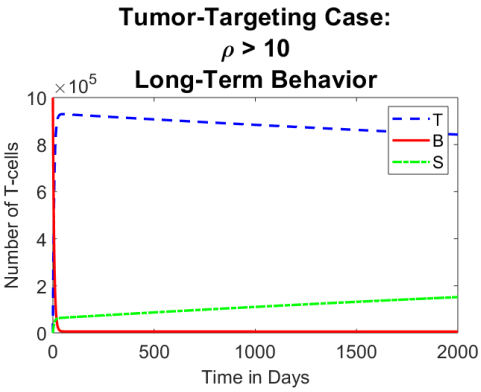
(a) Solutions from $t = 0$ to $t = 15$ when $\rho = 2$.



(b) Solutions from $t = 0$ to $t = 2000$ when $\rho = 2$.



(c) Solutions from $t = 0$ to $t = 15$ when $\rho = 15$.



(d) Solutions from $t = 0$ to $t = 15$ when $\rho = 2000$.

Figure 2. Numerical short-term and long-term solutions for our model, where the state variables, T, B, and S are the number of T-cells in the tumor, blood, and spleen, respectively. The other parameter values are fixed at values from Table 1. We explore both the Tumor-Ignoring cases (when $\rho \leq 10$) in (a) and (b) and Tumor-Targeting cases (when $\rho > 10$) in (c) and (d).

$\rho \approx 10^1$ while $\beta_{SB} \approx 10^{-4}$), it is useful to observe the normalized sensitivities,

$$\chi_n(\theta, t) = \frac{\partial f(t;\theta)}{\partial \theta} \frac{\theta}{f(t;\theta)}. \tag{2.8}$$

While general sensitivities, χ look at how the data reacts to the parameters as a whole by taking the partial derivative of the output factor with respect to the input factor, normalized sensitivities, χ_n, are scaled down by dividing the derivative by the observation in order to compare each parameter against the other.

The sensitivities and normalized sensitivities depend on both time t and the value of the parameter to which we calculate the observation's sensitivity, θ. Thus, for the calculation of sensitivities all parameters are fixed at the values listed in Table 1, and time t is allowed to vary. Since all of the parameter values are fixed, this is local (as opposed to global) sensitivity analysis. We use complex step method [25] to numerically evaluate the sensitivities in (2.7) and (2.8).

For all four categories (UT, CAR, CAR+CXCR1, and CAR+CXCR2) the concentration of T-cells in the tumor, blood, and spleen, are most sensitive to parameters βExt and ρ. Since graphs of the sensitivities look very similar for the different treatment categories, we only show graphs of the sensitivities for the last treatment, CAR+CXCR2, where $\rho = 30$. In order to better compare the sensitivities to each parameter, for each observation (T, B, and S) and parameter (see Table 1) we find the maximum sensitivity and maximum normalized sensitivity over time and plot these results in Figure 3. The full time-dependent sensitivities are plotted in Figures 4 and 5. Since the observations T, B, and S are consistently not sensitive to the initial conditions, sensitivities to T_0, B_0, and S_0 are not plotted in Figures 4 and 5.

The observations of T-cells in the tumor, blood, and spleen in Figure 3a are most sensitive to the T-cell movement rates, β_{Ext}, β_{TB}, and β_{SB}, followed by antigen recognition in the tumor, ρ. T-cell counts are not very sensitive to the initial conditions, B_0, T_0, and S_0. These results are consistent with our model design. When sensitivities are normalized, the T-cell counts in Figure 3b are most sensitive to the rate at which T-cell exit the blood, β_{Ext}, and antigen recognition in the tumor, ρ followed by the initial T-cell count in the blood, B_0. Thus, according to the normalized sensitivities, T-cell counts are most effected by parameters β_{Ext} and ρ. Since ρ is the transient expansion factor of antigen recognition in the tumor and changes depending on the treatment, we choose to estimate this parameter. We also notice that these sensitivities are time-dependent. The normalized sensitivities in Figure 5 shows that the observations are most sensitive to ρ and β_{Ext} initially, while they are sensitive to β_{TB} and β_{SB} later in time. This is an indication of the behavior of the ODE. Indeed, as T-cells are shuttled out of the blood (β_{Ext} via ρ) initially, they do so at a very dramatic rate, and the number of T-cells in the tumor, blood, and spleen are sensitive to those rates.

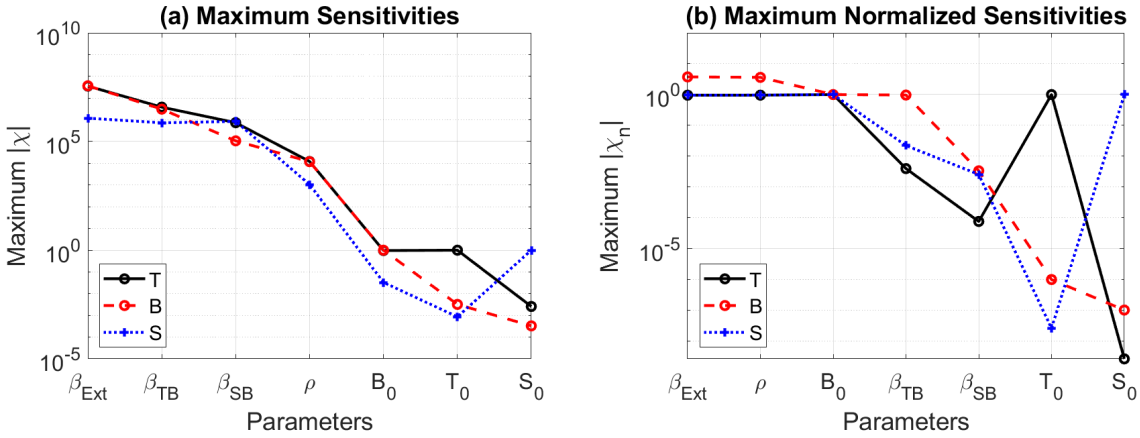


Figure 3. Maximum sensitivities (a) and maximum normalized sensitivities (b) of each observation, T, B, and S, to each of the parameters over a time period of 30 days. The expansion factor of T-cells in the tumor $\rho = 30$, and the initial number T-cells in the tumor, blood, and spleen $[T_0, B_0, S_0] = [0, 10^6, 0]$, since this is data from the CAR+CXCR2 treatment group.

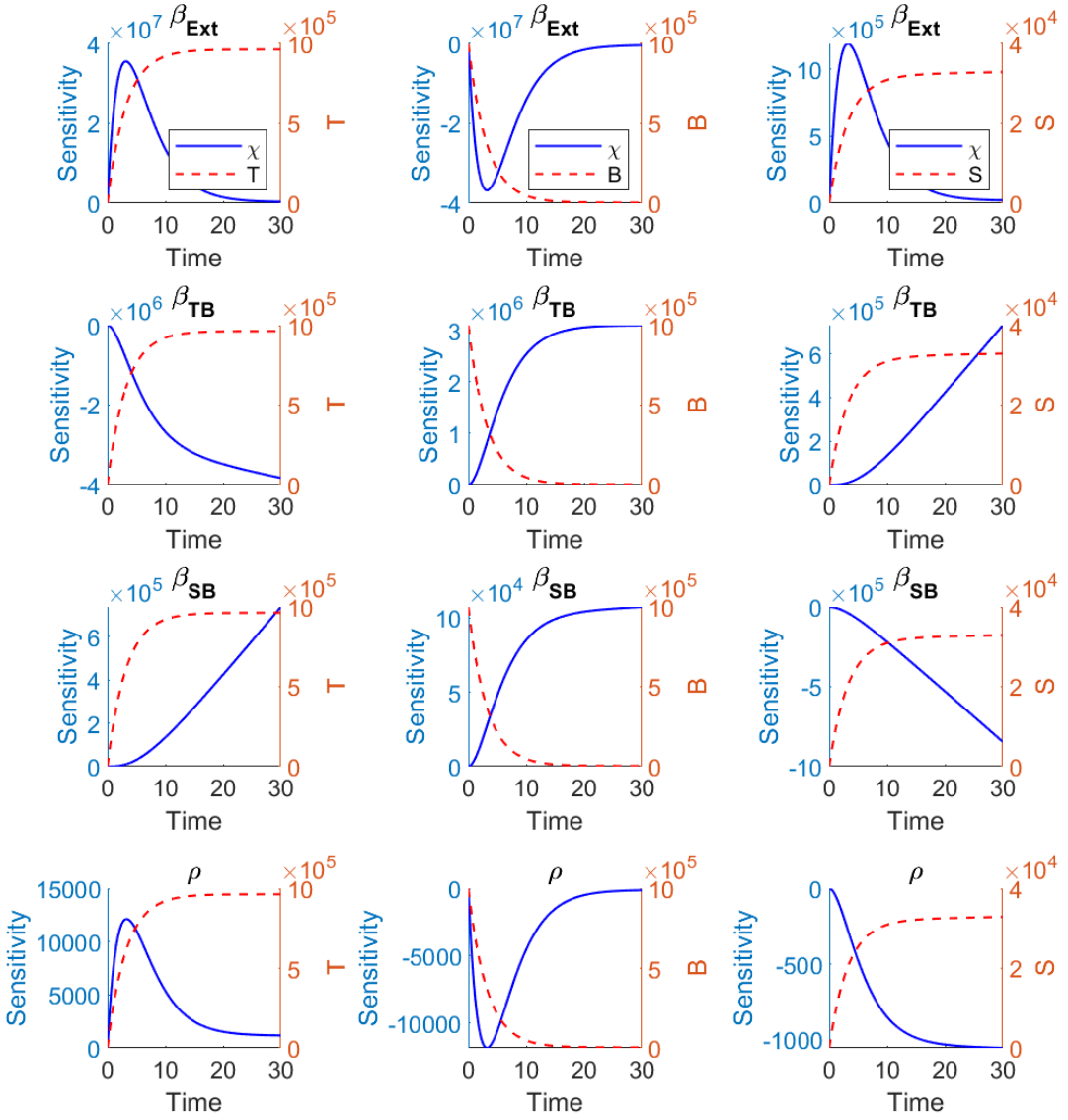


Figure 4. Sensitivities, $\chi(t)$, of observation T, B, and S to parameters β_{Ext}, β_{TB}, β_{SB}, and ρ over a time period of 30 days. These sensitivities are calculated at parameter values from Table 1 with $\rho = 30$ to represent the CAR+CXCR2 treatment.

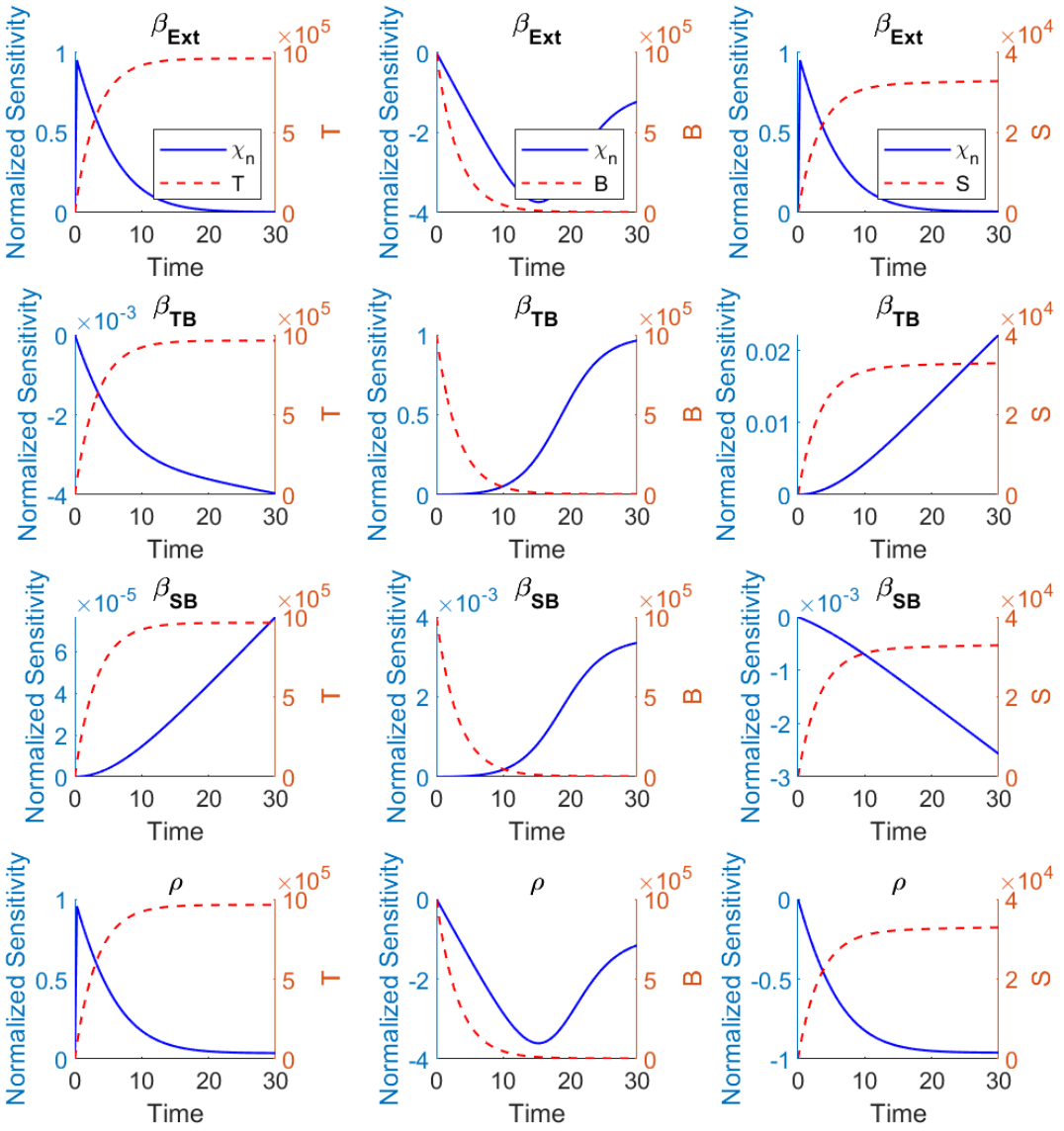


Figure 5. Normalized sensitivities, $\chi_n(t)$, of observation T, B, and S to parameters β_{Ext}, β_{TB}, β_{SB}, and ρ over a time period of 30 days. These sensitivities are calculated at parameter values from Table 1 with $\rho = 30$ to represent the CAR+CXCR2 treatment.

3. Aggregate model

3.1. Parameter estimation methodology

Equations (2.1), (2.2), and (2.3) model the T-cell counts in an individual mouse, which is assumed to have a single value for T-cell flow from the tumor to the blood, from the blood to the spleen, etc. (i.e., a single value for each of the parameters listed in Table 1). However, this assumption does not

apply to the aggregate data set, in which different mice sacrificed and sampled at each time point may have different parameter values (e.g., ρ, β_{Ext}, etc.), varying over some range of values. Thus, using the individual model we formulate an *aggregate* model with which we can compare the aggregate data (see Chapter 5 of [2] for a more complete discussion). Based on the sensitivity analysis and our interest in the different types of treatment, we choose to estimate the probability distribution of ρ, the transient expansion of T-cells in the tumor due to antigen recognition.

Consider the deterministic T-cell population vector $\boldsymbol{x}(t;\rho) = [T(t;\rho), B(t;\rho), S(t;\rho)]^T$ which is a solution to Eqs (2.1), (2.2), and (2.3) given parameter values in Table 1, where t is time and ρ is the transient expansion factor which we choose to estimate (see Section 2.3). There is an aggregate population vector $\boldsymbol{u}(t;p) = [u_T(t;P), u_B(t;P), u_S(t;P)]^T$ corresponding to the individual population vector $\boldsymbol{x}$ and given by

$$\boldsymbol{u}(t;P) = \mathbb{E}(\boldsymbol{x}(t;\cdot)|P) = \int_{\mathcal{G}} \boldsymbol{x}(t;\rho)dP(\rho),$$

where ρ is now a random variable, $\mathcal{G}$ is the collection of admissible parameter values for ρ, and P is a probability measure on $\mathcal{G}$. Note that $\boldsymbol{u}$ is the expected value of $\boldsymbol{x}$, which is also a random vector since it depends on the random variable ρ. Under the assumption that the probability distribution, P, possesses a probability density, p, and assuming that $\mathcal{G} = [\rho_l, \rho_u]$ is some closed interval, the population count is given by

$$\boldsymbol{u}(t;P) = \int_{\rho_l}^{\rho_u} \boldsymbol{x}(t;\rho)p(\rho)d\rho, \tag{3.1}$$

where the density $P' = \frac{dP}{d\rho} = p(\rho)$.

Now that we have an aggregate model which can be compared to the data, we follow techniques from Banks, Hu, and Thompson [2] and Banks, Bekele-Maxwell, Everett, Stephenson, Shao, and Morgenstern [26] to estimate parameters in our mathematical model. For the inverse problem, consider the 3-dimensional dynamical system, defined in (3.1) to be estimated using the data. We are interested in determining the probability density $P' = p(\rho)$ which gives the best fit of the underlying model to the aggregate data. However, this parameter estimation problem involves an infinite dimensional parameter space (the space $\mathcal{P}(\mathcal{G})$ of probability measures defined on the set $\mathcal{G}$). Instead of using a specific probability density function in the aggregate model, we use a family of finite approximations $\mathcal{P}^M(\mathcal{G})$. Based on [14, 27, 28], we are guaranteed convergence in the Prohorov metric. In our case the finite dimensional approximation $\mathcal{P}^M(\mathcal{G})$ to the probability measure space $\mathcal{P}(\mathcal{G})$ is defined using M linear splines.

Let $\boldsymbol{u}_j = [T_j, B_j, S_j]^T$ represent the T-cell count data or observations collected from the mice at time t_j for $j = 1, ..., N$. We note that there is some uncertainty between the actual phenomenon, which is represented through the data, and in the above observation process. This uncertainty is accounted for in the *statistical model* (again, see Chapter 3 of [2] where it is explained that both a *mathematical* model and a *statistical* model are required to carry out an inverse problem properly with uncertainty in observations)

$$\underbrace{\boldsymbol{U}_j}_{\text{data representation}} = \underbrace{\boldsymbol{u}(t_j;P_0^M)}_{\text{aggregate model}} + \underbrace{\boldsymbol{u}^{\boldsymbol{\gamma}}(t_j;P_0^M)\circ\boldsymbol{\mathcal{E}}_j}_{\text{weighted error}},$$

where P_A^M is the nominal probability density approximation. Note that P_0^M has a density $P^{M'} = \frac{dP^M}{d\rho} = p^M(\rho)$ which is defined using linear splines. The values $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3]^T, \gamma_i \geq 0$ in the error term

are weighting factors corresponding to each of the three aggregate observations (T-cells in the tumor, blood, and spleen) respectively. These weighting factors can be calculated from real data before an inverse problem is completed [29], but since this manuscript only deals with simulated data, these techniques are not discussed. These represent the dependency of error on the dynamics, and the model itself corresponds to the choices of data.

Note that $\circ$ is the Hammond or Schur product of component wise multiplication of two vectors. The random error vectors $\boldsymbol{\mathcal{E}}_j = [\mathcal{E}_j^1, \mathcal{E}_j^2, \mathcal{E}_j^3]^T$ corresponding to each of the three aggregate observations (T-cells in the tumor, blood, and spleen) respectively are assumed to be independent and identically distributed (*i.i.d.*) with mean zero, $\mathrm{Var}(\mathcal{E}_j^1) = \sigma_{01}^2$, $\mathrm{Var}(\mathcal{E}_j^2) = \sigma_{02}^2$, and $\mathrm{Var}(\mathcal{E}_j^3) = \sigma_{03}^2$. The corresponding realizations (for the random vector $\boldsymbol{U}_j$) are

$$\underbrace{\boldsymbol{u}_j}_{\text{aggregate data}} = \underbrace{\boldsymbol{u}(t_j; P_0^M)}_{\text{aggregate model}} + \underbrace{\boldsymbol{u}^{\boldsymbol{\gamma}}(t_j; P_0^M) \circ \boldsymbol{\epsilon}_j}_{\text{weighted error}}.$$

This multiplicative structure of the observational error in the above statistical model exists, because often in biological models the size of the resulting observation error is proportional to the size of the observations. A rather thorough discussion of these issues, along with concrete examples, is given in Chapter 3 of [2]. For $\boldsymbol{\gamma} \geq 0$, a generalized least squares method or an iterative reweighted weighted least squares (IRWLS) method [26] is appropriate to perform the inverse problem. In order to estimate $\hat{P}^M \approx P_0^M$ (or the corresponding density $\hat{p}^M(\rho) = p^M(\rho)$), we want to minimize the distance between the collected data and aggregate mathematical model, where the observables are weighted according to their variability and, for each observable, the observations over time are weighted unequally (once again, we refer the reader to [26] and [2] for a more detailed discussion and relevant examples). A detailed description of the IRWLS method applied to an aggregate model as described above is outlined in section 4.1 of [30].

If we assume $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3] = [0, 0, 0]$, then our statistical model is called an *absolute error model* and an ordinary least squares method is appropriate for parameter estimation. However, we believe it is more biologically realistic to assume the observation error is *proportional to the size of the observed quantity* i.e., a *relative error model* for our data sets and models investigated here. In the next section, we simulate aggregate data to test our aggregate model and the inverse problem.

3.2. Data simulation

Now that we have ascertained some specific features of the behavior of the mathematical model and its parameters, we want to carry out a series of inverse problems to estimate the approximate probability distribution P_0^M (or its corresponding density $p^M(\rho)$) of the desired parameter ρ for a given number of time observations. We attempt to estimate P_0^M using observations of the aggregate engineered T-cell concentrations in the tumor T, blood B, and spleen S compartments. However, since currently available data sets contain data at only three distinct time observations, our inverse problem might not be feasible. Nonetheless we proceed in our efforts using a rather straightforward if unsophisticated approach to the question of how many mice must be sacrificed in order to reliably estimate a finite approximation P_0^M to the desired parameter distribution of ρ.

In order to see this problem from a bigger picture and determine how many time points are needed to accurately estimate parameters, we now step away from the experimental data and simulate our own data. We wish to mirror the experiment the best way possible. That is, our simulation will assume

aggregate data per time point, and we will focus upon a parameter estimation for the CAR treatment, which assumes an parameter value of $\rho \approx 15$ for the individual model. Thus, we will initialize our problem with an expected or assumed distribution of ρ, where $10 \leq \rho \leq 20$.

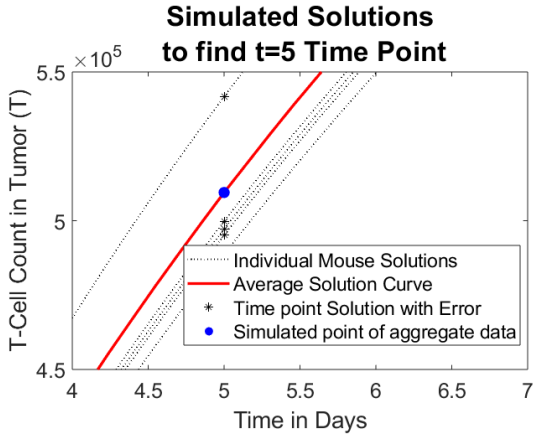
Our methods of data simulation are as follows: We first generate a normal distribution for ρ, where $\rho \sim \mathcal{N}(15, 1)$, that is, we assume that ρ has a mean value of $\mu = 15$ and has a standard deviation of $\sigma = 1$ to account for the differences in the mice. We can call this "actual" distribution P_A and the corresponding probability density function $p_A(\rho)$. Since our actual experimental data comes from a time frame of 0 to 15 days, we simulate data based on this same time frame and assume the same initial conditions and parameter values from Table 1. Although it may be advantageous to sample more frequently at certain times in the experiment, for simplicity we only generate uniformly spaced time points. In order to simulate the sacrifice of five mice per time point as in the experiments, we generate five data points from the model for each observation time point. To do this, we use the distribution P_A, from which we randomly draw out one value of ρ for each mouse. Thus, if we want to simulate n time points of data, we will draw $n \times 5$ values of ρ_{ij} for $i = 1, ..., 5$ mice per $j = 1, ..., n$ time points. This generates a set of five data points per time observation to which we add noise.

Since we are using the iterative reweighted weighted least squares approach described in the previous section, we set the variance of the random error vectors to be $\text{Var}(\mathcal{E}_j^1) = \text{Var}(\mathcal{E}_j^2) = \text{Var}(\mathcal{E}_j^3) = 0.01$, and we set the weighting factors to be $\boldsymbol{\gamma} = [0.5, 0.5, 0.5]^T$. We then average these simulated data values to give us one aggregate data point per time point for each of the observations T, B, and S. Thus, our simulated data takes the form

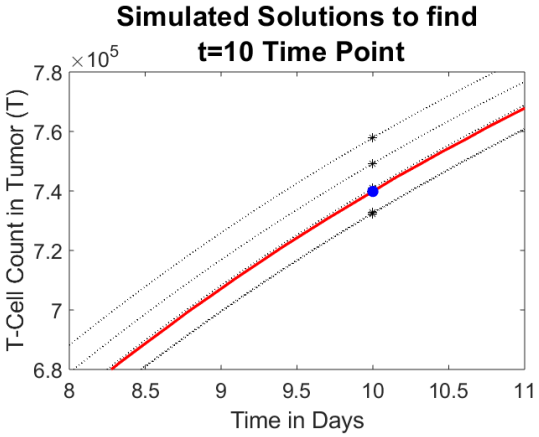
$$\boldsymbol{u}_j = [T_j, B_j, S_j]^T = \frac{1}{5}\sum_{i=1}^{5}[\boldsymbol{x}(t_j; \rho_{ij}) + \boldsymbol{x}^{\gamma}(t_j; \rho_{ij})\boldsymbol{\epsilon}_{ij}] \tag{3.2}$$

where $\boldsymbol{\epsilon}_{ij}$ are realizations of the normally distributed random vector $\boldsymbol{\mathcal{E}}_{ij}$ for each of the $i = 1, ..., 5$ mice at each of the $j = 1, ..., n$ time points.

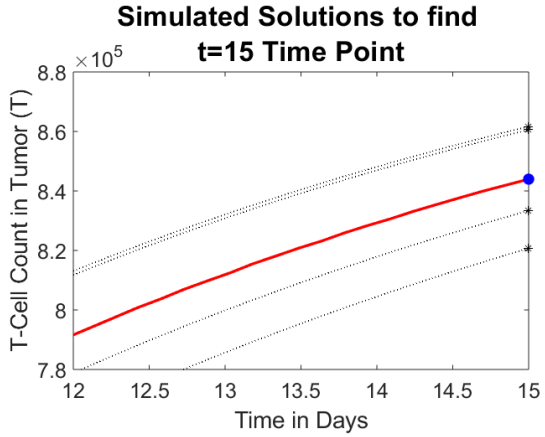
A qualitative image of this simulation for a simple time grid of $n = 4$ (or $n = 3$ data points, since we do not consider the first point at $t = 0$ to be a data point), for T-cells in the tumor only, can be seen in Figure 6a–c. For the first time point, $t = 5$, 5 mice are simulated, the T-cell counts in their tumors are plotted in Figure 6a, and each of these counts have weighted noise added to them. The same is done for the second ($t = 10$ in Figure 6b) and third time points ($t = 15$ in Figure 6c) for a total of 15 simulated mice each with a different value of ρ and added weighted noise. For each time point, we also plot the average solutions (or average T-cell counts in the tumors of the five mice), and the average noisy T-cell count in the tumor, which forms the simulated data points T_j as described in (3.2). In Figure 6d the density of the "true" distribution of ρ, $p_A(\rho)$ and the corresponding 15 randomly chosen ρ_{ij} realizations (for $i = 1, ..., 5$ mice per $j = 1, ..., n$ time points, t, after $t = 0$) to generate our simulated data. Using this simulated data, we investigate the results of the inverse problems assuming availability of different numbers of equally spaced time point observations of the aggregate T-cell concentrations in the tumor, blood, and spleen, and attempt to answer the question: how many time points are necessary for accurate parameter estimation?



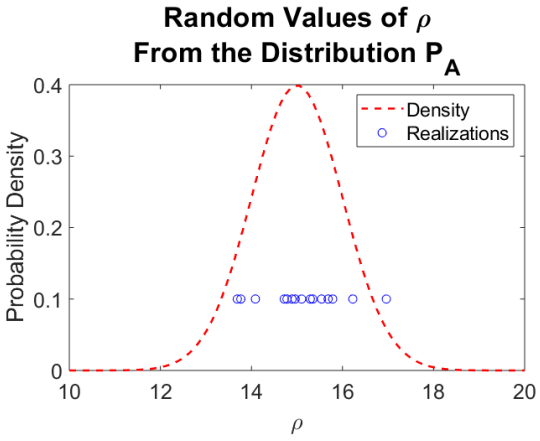
(a) Five simulated data points, averaged, from five different solutions, for $t = 5$.



(b) Five simulated data points, averaged, from five different solutions, for $t = 10$.



(c) Five simulated data points, averaged, from five different solutions, for $t = 15$.



(d) Randomly generated ρ values from a uniform distribution, P_A, from which solutions were found to simulate aggregate data.

Figure 6. For $n = 4$ time points, data is simulated from different ρ values per mouse per time point (dotted lines), with an initial time point of $t = 0$, and parameter values and initial conditions from Table 1. We add noise to the solution found at each time point (stars). These solutions are averaged (solid line) to show one single data point per time observation (big dot). For the sake of simplicity and illustration, these solutions show the T-cells in the tumor only.

3.3. Error quantification

Before performing inverse problems, we establish the methodology used to determine how accurate an estimated probability distribution of ρ is using a given simulated data set. The estimated distribution $\hat{P}^M$ (and its corresponding density $\hat{p}^M$ will most likely differ from the "actual" distribution P_A (and its corresponding density p_A). While we can visually compared these two probability density functions, it is useful to quantify their differences. Since the density function $\hat{p}^M$ is defined using M linear spline functions, this function will have $k = 0, ..., M$ spline nodes ρ_k. Thus, we use the L^2 norm to calculate

the difference between $\hat{p}^M$ and the "actual" density p_A at each of these nodes by measuring the sum of squared differences between the approximated spline nodes and their corresponding solution on the normal curve. Thus, this L^2 norm is defined by

$$||\Delta p||_2 = \sqrt{\sum_{k=0}^{M} \left| p_A(\rho_k) - \hat{p}^M(\rho_k) \right|^2} \tag{3.3}$$

where $p_A(\rho)$ is the true PDF (probability density function) of P_A, and $\hat{p}^M(\rho)$ is the estimated spline approximation of the true PDF, calculated for each ρ_k nodes, $k = 0, ..., M$. Similarly, the infinity norm, $||\Delta p||_\infty$, which takes the maximum vectored error, is defined by

$$||\Delta p||_\infty = \max(|\Delta p_0|, |\Delta p_1|, ..., |\Delta p_M|) \tag{3.4}$$

where $\Delta p = (p_A(\rho_k) - \hat{p}^M(\rho_k))$ for each ρ_k node, $k = 0, ..., M$.

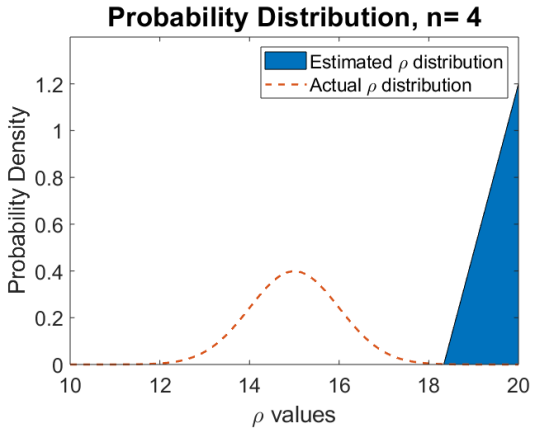
4. Results

4.1. Case 1: $n = 4$ time points

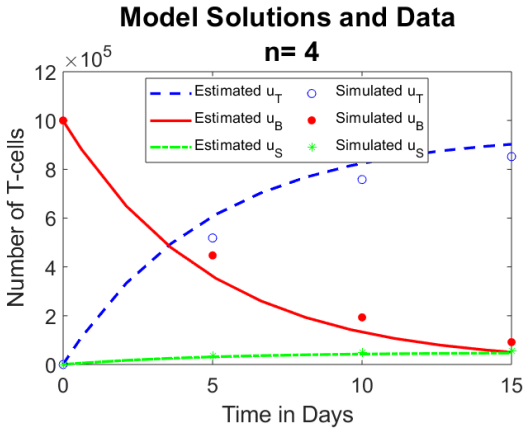
In Case 1, we simulate as few as 4 time points and attempt to estimate the probability distribution. Figure 7a shows the results of the estimated distribution from the inverse problem graphed against the "actual" distribution of ρ, which is $\mathcal{N}(15, 1)$. As we see in Figure 7a, the estimated probability density of ρ, $\hat{p}^M$, which is shaded in, does not overlap with the "actual" probability density of ρ, p_A, which was previously assumed. (Note that the probability densities $\hat{p}^M$ and p_A each have a corresponding probability distributions $\hat{P}^M$ and P_A, respectively.) To quantify this difference, we can look at the L^2 norm, $||\Delta p||_2 = 1.27$. In Figure 7b we plot $n = 4$ time points of the simulated aggregate data, which were simulated under the assumption that $\rho \sim \mathcal{N}(15, 1)$. As can be seen, there is a significant amount of systematic (non-random) error between the simulated data and the estimated solution. Figure 7c plots the residual errors between the approximated solution and the simulated data points for each of the three state solutions, T, B, and S (tumor, blood, and spleen). We see they are fairly normally distributed and decrease per time point. Errors are quite high for the number of T-cells in the tumor and blood solutions, but quite small for the spleen, which is most likely because the flow of T-cells to and from the spleen does not rely on the ρ parameter.

This discrepancy in Figure 7a is most likely due to the fact that 4 time observations is too few to glean sufficient information. As such, we get a skewed distribution of the parameter ρ, which does not provide an accurate result as to the true dynamics of the system. Finally, we investigate the condition number of the Fisher Information Matrix (FIM). This condition number is used to determine the uncertainty in the estimated probability distribution and compare the uncertainty in different cases. The FIM is approximated using the sensitivity and covariance matrices, the method through which is described in [30]. It is known that if the condition number of the Fisher Information Matrix is very large, then there is more uncertainty regarding the probability density coefficient estimates. For case 1, the condition number of the FIM is 2.93×10^{17}. We note that we are using $n = 4$ data points to estimate $M = 6$ linear spline functions (which is defined using 7 spline nodes), so this is an ill-posed least squares problem (i.e., there are more unknown parameters than known data points), [31] which

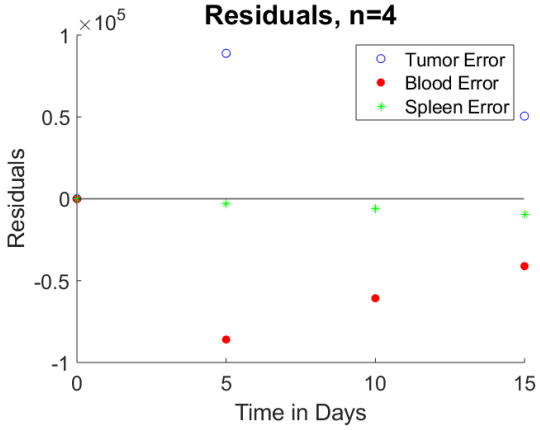
is unsolvable. Even if we attempt to use $M = 2$ spline functions (3 spline nodes), we are left with 0 degrees of freedom. Thus, more data points are needed.



(a) Estimated probability density $\hat{p}^M$ of ρ, compared to the "actual" normal probability density p_A of $\rho \sim \mathcal{N}(\mu = 15, \sigma = 1)$.



(b) $n = 4$ time points of simulated aggregate data compared to the aggregate solutions using the estimated probability distribution $\hat{P}^M$.

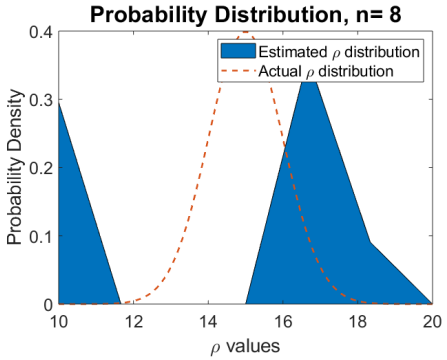


(c) Residuals $\boldsymbol{u}_j - \boldsymbol{u}(t_j; \hat{P}^M)$ for $n = 4$ time points, between the approximated solution and the simulated data.

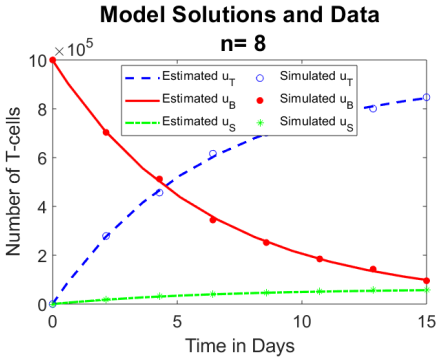
Figure 7. Case 1: $n = 4$ time points of simulated aggregate observations of the number of T-cells in the tumor u_T, blood u_B, and spleen u_S, (b), assuming that ρ is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 1$, used to estimate the probability distribution $\hat{P}^M$, (a). All parameter values except for ρ are set at values from Table 1.

4.2. *Case 2: $n = 8$ time points*

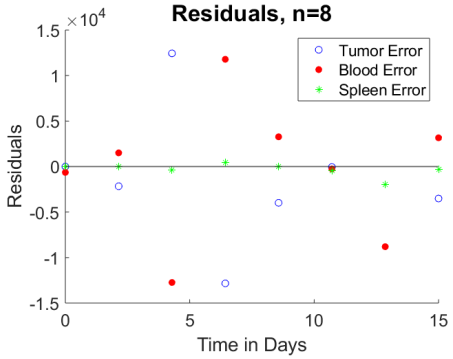
We now consider $n = 8$ time points of simulated aggregate data. In this case, the estimated distribution $\hat{P}^M$of ρ is bimodal, encompassing extreme sides of P_A, and still fails to capture the "actual" probability distribution that was assumed, as can be seen in Figure 8a. The L^2 norm in this case is $\|\Delta p\|_2 = 0.577$, however, which shows a significant decrease from when $n = 4$, meaning that the approximation is improving. Figure 8b shows considerable less error between the simulated points and the solution based on the estimated probability distribution for ρ. Figure 8c plots the residuals, and we see a lower order of magnitude for each error. The condition number of the FIM is 1.18×10^{17}, which is still the same magnitude as in Case 1.



(a) Estimated probability density $\hat{p}^M$ of ρ, compared to the "actual" normal probability density p_A of $\rho \sim \mathcal{N}(\mu = 15, \sigma = 1)$.



(b) $n = 8$ time points of simulated aggregate data, compared to the solution using the estimated probability distribution.

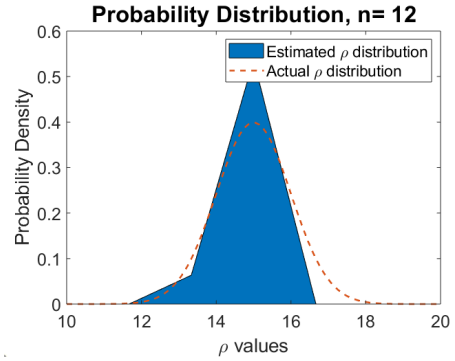


(c) Residuals $\boldsymbol{u}_j - \boldsymbol{u}(t_j; \hat{P}^M)$ for $n = 8$ time points, between the approximated solution and the simulated data

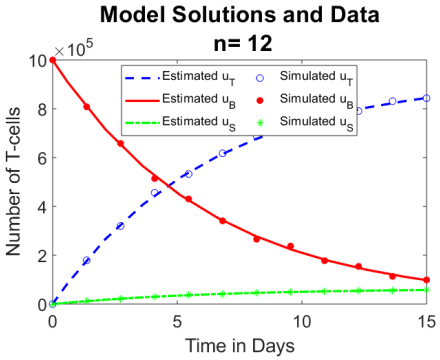
Figure 8. Case 2: $n = 8$ time points of simulated aggregate observations, assuming that ρ is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 1$, of the number of T-cells in the tumor T, blood B, and spleen S, used to estimate the probability distribution of ρ, and compared to estimated aggregate observations of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

4.3. *Case 3: $n = 12$ time points*

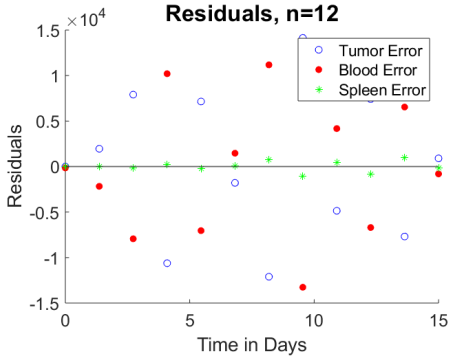
In Case 3, we slightly increase the number of time points to $n = 12$, and already see a dramatic increase in accuracy. Figure 9a demonstrates that the approximated probability density of ρ has overlapped with the "actual" density, with only a small deviation on the left, which can be quantified by the L^2 norm of $\|\Delta p\|_2 = 0.174$, which is even lower than previous cases. It can be seen in Figure 9b that the approximated solutions are matching the simulated data quite well. The residuals of the tumor, T, and blood, B, observations in Figure 9c are much larger than the residuals of the spleen observation S, because the T-cell counts in the blood and spleen are much larger. The FIM is 5.2×10^{16},which is one order magnitude smaller than previous cases.



(a) Estimated probability distribution of ρ, compared to the "actual" normal probability density p_A of $\rho \sim \mathcal{N}(\mu = 15, \sigma = 1)$.



(b) $n = 12$ time points of simulated aggregate data, compared to the solution using the estimated probability distribution.

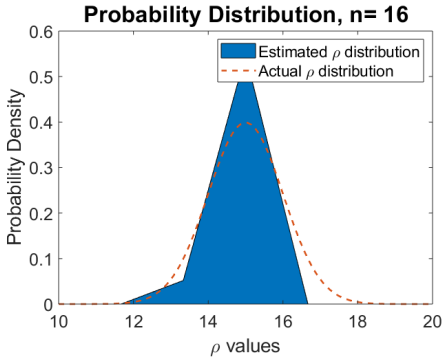


(c) Residuals $\boldsymbol{u}_j - \boldsymbol{u}(t_j; \hat{P}^M)$ for $n = 12$ time points, between the approximated solution and the simulated data

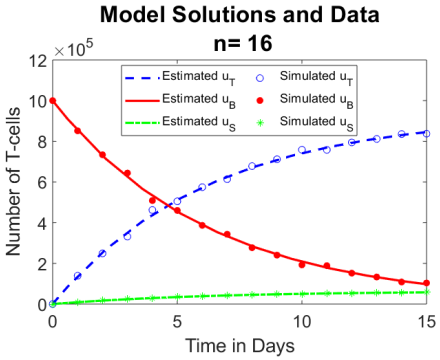
Figure 9. Case 3: $n = 12$ time points of simulated aggregate observations, assuming that ρ is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 1$, of the number of T-cells in the tumor T, blood B, and spleen S, used to estimate the probability distribution of ρ, and compared to estimated aggregate observations of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

4.4. *Case 4: $n = 16$ time points*

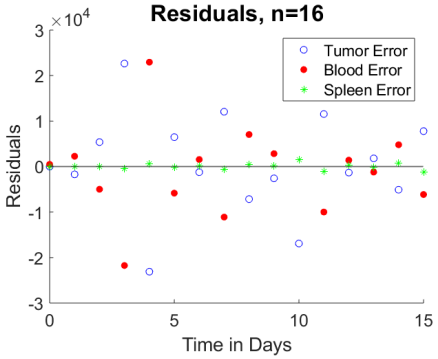
In Case 4, we increase the number of time points to 16. From Figure 10a, we see that the estimated probability density of ρ still matches the "actual" density very well, with a slight leftward difference. The L^2 norm is $\|\Delta p\|_2 = 0.186$, which is slightly larger than in Case 3. We still see a reasonable fit between the simulated data and the solution curve in Figure 10b, while Figure 10c shows that the residuals between the solution and data are similar in magnitude to residuals in Case 3. The condition number of the FIM is 3.45×10^{16}.



(a) Estimated probability density $\hat{p}^M$ of ρ, compared to the "actual" normal probability density p_A of $\rho \sim \mathcal{N}(\mu = 15, \sigma = 1)$.



(b) $n = 16$ time points of simulated aggregate data, compared to the solution using the estimated probability distribution

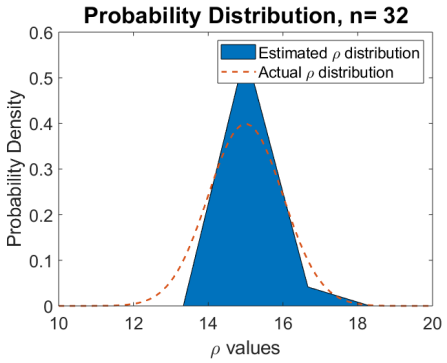


(c) Residuals $\boldsymbol{u}_j - \boldsymbol{u}(t_j; \hat{P}^M)$ for $n = 16$ time points, between the approximated solution and the simulated data

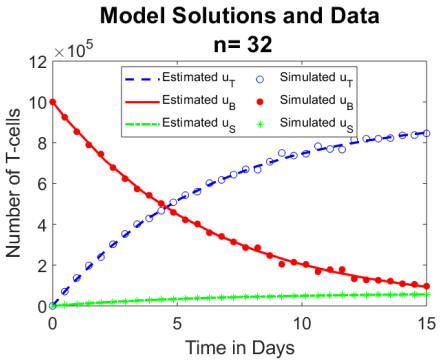
Figure 10. Case 4: $n = 16$ time points of simulated aggregate observations, assuming that ρ is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 1$, of the number of T-cells in the tumor T, blood B, and spleen S, used to estimate the probability distribution of ρ, and compared to estimated aggregate observations of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

4.5. *Case 5: $n = 32$ time points*

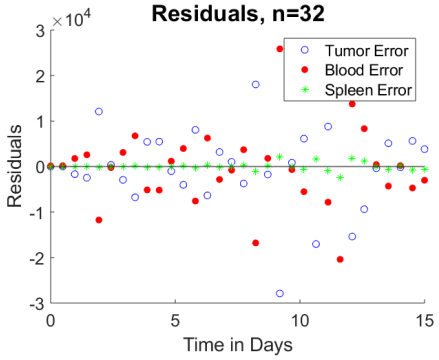
In Case 5, we double the number of simulated time points to $n = 32$. As in Case 4, the estimated probability density of ρ is aligned with the "actual" distribution, as seen in Figure 11a, with an L^2 norm of $\|\Delta p\|_2 = 0.197$, which i slightly larger than in Case 5. Thus, the approximation is becoming slightly worse as the number of time points increase. The condition number of the FIM is 1.67×10^{16}, which has a similar magnitude as Case 3 as well. Figure 11b shows that the approximated solution curve fits better with the simulated data as time points are increased. The residuals, plotted in Figure 11c, are also similar to the residuals plotted in Case 3.



(a) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 15, \sigma = 1)$.



(b) $n = 32$ time points of simulated aggregate data, compared to the solution using the estimated probability distribution

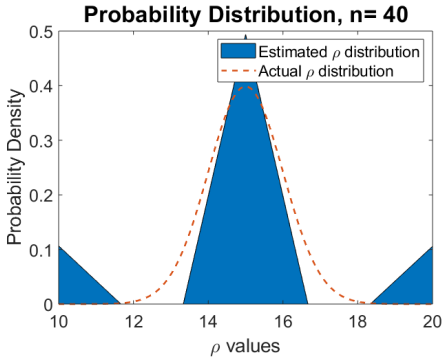


(c) Residuals $\boldsymbol{u}_j - \boldsymbol{u}(t_j; \hat{P}^M)$ for $n = 32$ time points, between the approximated solution and the simulated data

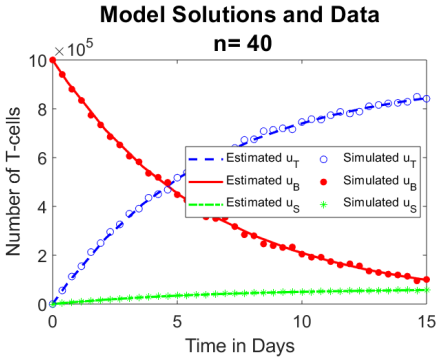
Figure 11. Case 5: $n = 32$ time points of simulated aggregate observations, assuming that ρ is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 1$, of the number of T-cells in the tumor T, blood B, and spleen S, used to estimate the probability distribution of ρ, and compared to estimated aggregate observations of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

4.6. *Case 6: $n = 40$ time points*

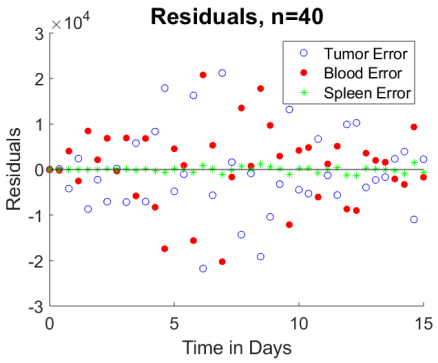
Now that we notice a pattern of convergence between the approximated distribution and the actual distribution to which we are comparing, with little change as we increase the time points from $n = 16$ to $n = 32$, we can assume that our approximation has reached its peak and can produce no better results. Indeed, in Case 6, we look at a very small increase in time points, for $n = 40$. Figure 12a, shows that the approximated probability distribution for ρ has multiple modes, and now has outliers on the far left and right, deviating further from the "actual" distribution. Indeed, the L^2 norm for the probability distribution, $\|\Delta p\|_2 = 0.227$, is larger than previous cases. Likewise, the condition number of the FIM is 1.45×10^{17}, which is an order of magnitude larger than that for our best n.



(a) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 15, \sigma = 1)$.



(b) $n = 40$ time points of simulated aggregate data, compared to the solution using the estimated probability distribution



(c) Residuals $\boldsymbol{u}_j - \boldsymbol{u}(t_j; \hat{P}^M)$ for $n = 40$ time points, between the approximated solution and the simulated data

Figure 12. Case 6: $n = 40$ time points of simulated aggregate observations, assuming that ρ is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 1$, of the number of T-cells in the tumor T, blood B, and spleen S, used to estimate the probability distribution of ρ, and compared to estimated aggregate observations of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

While Figures 12b and 12c show a good line fit, with random residuals, our estimated probability density is clearly inaccurate. This indicates that increasing the number of time points without bound will not consistently result in a better approximation. In fact, an n that is too large may actually provide an incorrect estimate. This phenomenon may be due to the fact that it is possible to run into trouble with spline-based methods for a very large number of data points. Indeed, this may produce an ill-posedness due to excessive computational error in the inverse problems, and is discussed more in [31, 32].

5. Conclusions and discussion

Sensitivity Analysis Takeaway: For all four categories (UT, CAR, CAR+CXCR1, and CAR+CXCR2) the model observations T, B, and S (the number of T-cells in the tumor, blood, and spleen, respectively) are most sensitive to parameters ρ and β_{Ext}. Because of this, we can consider these parameters the most important when comparing the data to the model.

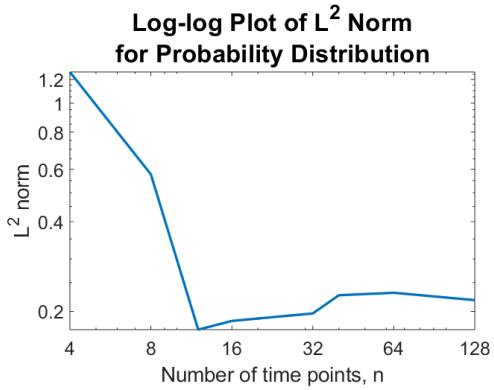
Stability Analysis Takeaway: At different values of ρ, the transient expansion factor of tumor antigen recognition by the immune system, the mathematical model described in (2.1)–(2.3) has different long-term behaviors. Behavior changes when $\rho = 10$. For $\rho \leq 10$, we see that the transient expansion factor is not enough and the T-cells essentially ignore the tumor in the long term. While T-cells in the blood decrease to zero, they do so slowly with a rise in the T-cells in the spleen. Biologically speaking, a situation in which $\rho \leq 10$ indicates that the body is not fighting the foreign object as it should, and the T-cells are not doing their job. For $\rho > 10$, we see that there is a large-enough transient expansion term for the T-cells to exit the blood and target the tumor, which is to be expected in a cancer treatment, and is thus biologically relevant. In this case, T-cells in the blood both decrease to zero and increase in the tumor very quickly. It should be noted that in our analysis, we look at both short and long term behavior. In actuality, our specific model will only consider a time-line of, at most, a few years (several hundred days). However, it is important to understand long term behavior of the system.

Parameter Estimation Takeaway: In order to save on costly experiments, we should use the minimum number of data points necessary to feasibly estimate the probability distribution of our parameter of interest, ρ. Utilizing our estimation methods and the aggregate version of our model with fixed parameters set at biologically relevant values (see Table 1), we find that between $n = 12$ time points results in the most accurate estimated distribution of ρ. With $n < 12$ time points, we have inaccurate results, and with $n > 32$, not only do our results not improve, but they become less accurate.

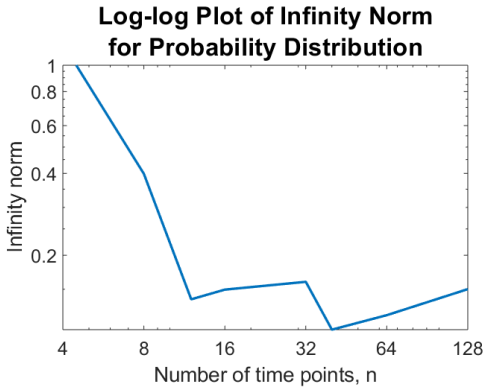
Since we are simulating data, it is possible to compare the estimated and "actual" probability distributions. By investigating norms of these differences, we can see how the number of time points, n, influences the error between the true probability density function and the approximated spline estimation. Figure 13a shows that for smaller values of n, the error between the true PDF of P_A, $p_A(\rho)$, and the approximated spline estimation, $\hat{p}^M(\rho)$ is highest, and decreases the most at $n = 12$. It then slowly starts to rise again, but not dramatically. Still, though, the L^2 norm only increases as n increases for even the largest values, reinforcing the fact that the ideal approximation for this problem does depend on an ideal n. Figure 13b shows the result of the infinity norms as we increase our number of time points taken. Again, we see that the maximum errors are highest for small values of n, and decrease for $n = 12$ and $n = 40$. We do know that the approximation is not ideal for $n = 40$, from the L^2 norm, but its maximum norm is indeed smallest.

Finally, we see from Figure 13c that the condition number of the Fisher Information Matrix (FIM)

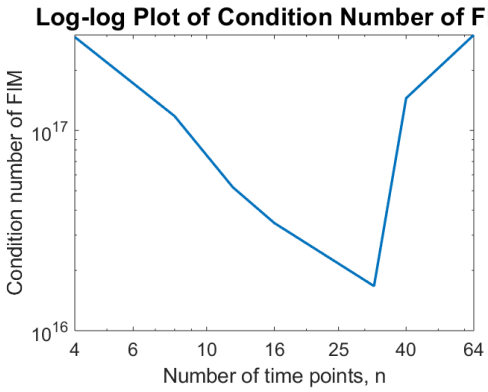
for the approximate solutions is highest at both very small and large values of n. As we approach the ideal number of time points for an accurate approximation, the condition number dips down. We can see that for all solutions, the magnitude of the condition number for the FIM is $10^{16} - 10^{17}$, which is somewhat high. While there is actually no ideal value for the condition number of the FIM, a smaller number indicates that the uncertainty factor in the approximation is improving. Regardless, it is a useful tool to compare conditions of different inverse problems, in a relative rather than absolute fashion.



(a) L^2 norm for the probability distribution as we increase our number of n observations



(b) Infinity norm for the probability distribution as we increase our number of n observations



(c) The condition number of the Fisher Information Matrix of the approximate solution as the number of time points, n is changed

Figure 13. A comparison of errors and accuracy between the approximated spline estimation solution for the parameter ρ, and the true probability density function of P_A. We use both the L^2 norm and the infinity norm, as well as the condition number for the Fisher Information Matrix. All results are based on approximated and simulated data, assuming the CAR treatment in which we assume $10 \leq \rho \leq 20$.

Although these methods lead to satisfactory results that inform future data collection, our method of experimental design is still very elementary. (For example, the aggregate time points are assumed to be equally spaced, which limits the possibilities for experimental design.) None of these efforts

involve use of the sophisticated optimal experimental design formulations outlined in Section 6 below. An obvious next step would include use of these design ideas to attempt to further refine the number of and specific times of the needed observations to successfully carry out the needed distributional estimations with aggregate data.

6. Future work

We turn to the question of how best to design experiments to collect data (how much data? and when to collect it?) necessary to validate models with only aggregate data available. To this point we have discussed various aspects of uncertainty arising in inverse problem techniques. All discussions have been in the context of *a given set* or *sets* of data carried out under various assumptions on how (e.g., independent sampling, absolute measurement error, relative measurement error) the data were collected. For many years now [33–40] scientists (and especially engineers) have been actively involved in designing experimental protocols to best study engineering systems, including parameter-describing mechanisms. Recently, with increased involvement of scientists working in collaborative efforts with ecologists, biologists, and quantitative life scientists, renewed interest in design of "best" experiments to elucidate mechanisms has been seen [33]. Thus, a major question that experimentalists and inverse problem investigators alike often face is how best to collect the data to enable one to efficiently and accurately estimate model parameters. This is the well-known and widely studied *optimal design* problem. A rather through review is given in [2]. Briefly, traditional optimal design methods (D-optimal, E-optimal, c-optimal) [34–37] use information from the model to find the sampling distribution or mesh for the observation times (and/or locations in spatially distributed problems) that minimizes a design criterion, quite often a function of the Fisher Information Matrix (FIM). Experimental data taken on this optimal mesh are then expected to result in accurate parameter estimates. We briefly mention a framework based on the FIM for a system of ordinary differential equations (ODEs) to determine *when an experimenter should take samples* and *what variables to measure* when collecting information on a physical or biological process modeled by a dynamical system.

Inverse problem methodologies are often discussed in the context of a dynamical system or mathematical model where a sufficient number of observations of one or more states (variables) are available. The choice of method depends on assumptions the modeler makes on the form of the error between the model and the observations (the statistical model). The most prevalent source of error is observation error, which is made when collecting data. (One can also consider model error, which originates from the differences between the model and the underlying process that the model describes. However, this is often quite difficult to quantify.) Measurement error is most readily discussed in the context of statistical models. The three techniques commonly addressed are *maximum likelihood estimation (MLE)*, used when the probability distribution form of the error is known; *ordinary least squares (OLS)*, for error with constant variance across observations; and *generalized least squares (GLS)*, used when the variance of the data can be expressed as a non-constant function. Uncertainty quantification is also described for optimization problems of this type, namely in the form of observation error covariances, standard errors, residual plots, and sensitivity matrices. Techniques to approximate the variance of the error are also included in these discussions. In [41], the authors develop an experimental design theory using the FIM to identify optimal sampling times for experiments on physical processes (modeled by an ODE system) in which scalar or vector data is taken.

In addition to when to take samples, the question of what variables to measure is also very important in designing effective experiments, especially when the number of state variables is large. Use of such a methodology to optimize what to measure would further reduce testing costs by eliminating extra experiments to measure variables neglected in previous trials [42]. In the CAR-T therapy example presented in this paper, it may only be necessary to measure, for example, the T-cells in the blood and the tumor or the tumor and the spleen, etc in order to obtain accurate estimations of ρ. In [43], the best set of variables for an ODE system modeling the Calvin cycle is identified using two methods. The first, an ad-hoc statistical method, determines which variables directly influence an output of interest at any one particular time. Such a method does not utilize the information on the underlying time-varying processes given by the dynamical system model. The second method is based on optimal design ideas. Extension of this method is developed in [44, 45]. Specifically, in [44] the authors compare the SE-optimal design introduced in [46] and [41] with the well-known methods of D-optimal and E-optimal design on a six-compartment HIV model [47] and a thirty-one dimensional model of the Calvin Cycle. Such models, in which a wide range of possible observational variables exist, are not only ideal through which to test the proposed methodology, but are also widely encountered in applications.

Acknowledgments

This research was supported in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-18-1-0457.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. T. Banks, Z. R. Kenz and W. C. Thompson, A review of selected techniques in inverse problem nonparametric probability distribution estimation, CRSC-TR12-13, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, May 2012; *J. Inverse Ill-Pose. Probl.*, **20** (2012), 429–460.
2. H. T. Banks, S. Hu and W. C. Thompson, Chapter 5 of *Modeling and Inverse Problems in the Presence of Uncertainty,* Chapman and Hall/CRC, New York, 2014.
3. H. T. Banks, L. W.Botsford, F. Kappel, et al., Modeling and estimation in size structured population models, LCDS/CCS Rep. 87-13, March, 1987, Brown Univ.; *Proc. 2nd Course on Math. Ecology* (Trieste, December, 1986), *World Scientific Press*, Singapore (1988), 521–541.
4. H. T. Banks and H. T. Tran, Chapter 9.7 of *Mathematical and Experimental Modeling of Physical and Biological Processes, CRC Press*, Boca Raton, FL, January 2, 2009
5. H. T. Banks, J. L. Davis, S. L. Ernstberger, et al., Experimental design and estimation of growth rate distributions in size-structured shrimp populations, CRSC TR08-20, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, November, 2008; *Inverse Probl.*, **25** (2009), 095003 (28 pp).

6. H. T. Banks, L. W. Botsford, F. Kappel, et al., Estimation of growth and survival in size-structured cohort data: An application to larval striped bass (*Morone saxatilis*), CAMS Tech. Rep. 89-10, University of Southern California, 1989; *J. Math. Biol.*, **30** (1991), 125–150.
7. H. T. Banks and B. G. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, CAMS Tech. Rep. 90-2, University of Southern California, January, 1990; *Q. Appl. Math.*, **49** (1991), 215–235.
8. D. Abate-Daga and M. L. Davila, CAR models: next-generation CAR modifications for enhanced T-cell function, *Mol. Ther-Oncolytics*, **3** (2016), 16014.
9. E. K. Moon, C. Carpenito, J. Sun, et al., Expression of a functional CCR2 receptor enhances tumor localization and tumor eradication by retargeted human T cells expressing a mesothelin-specific chimeric antibody receptor, *Clin. Cancer Res.*, **17** (2011), 4719–4730.
10. H. T. Banks and N. L. Gibson, Well-posedness in Maxwell systems with distributions of polarization relaxation parameters, CRSC-TR04-01, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, January, 2004; *Appl. Math. Lett.*, **18** (2005), 423–430.
11. H. T. Banks and N. L. Gibson, Electromagnetic inverse problems involving distributions of dielectric mechanisms and parameters, CRSC-TR05-29, August, 2005; *Q. Appl. Math.*, **64** (2006), 749–795.
12. H. T. Banks and D. M. Bortz, Inverse problems for a class of measure dependent dynamical systems, *J. Inverse Ill-pose. Probl.*, **13** (2005), 103–121.
13. H. T. Banks, D. M. Bortz and S. E. Holte, Incorporation of variability into the mathematical modeling of viral delays in HIV infection dynamics, *Math. Biosci.*, **183** (2003), 63–91.
14. H. T. Banks, D. M. Bortz, G. A. Pinter, et al., Modeling and imaging techniques with potential for application in bioterrorism, CRSC TR03-02, January 2003; Chapter 6 in Bioterrorism: Mathematical Modeling Applications in Homeland Security (H.T. Banks and C. Castillo-Chavez, eds.), *Front. Appl. Math.*, FR28, SIAM, Philadelphia, 2003, 129–154.
15. H. T. Banks, J. H. Barnes, A. Eberhardt, et al., Modeling and computation of propagating waves from coronary stenosis, *Comput. Appl. Math.*, **21** (2002), 767–788.
16. H. T. Banks, S. Hu, Z. R. Kenz, et al., Material parameter estimation and hypothesis testing on a 1D viscoelastic stenosis model: methodology, CRSC-TR12-09, April, 2012; *J. Inverse Ill-pose. Probl.*, **21** (2013), 25–57.
17. H. T. Banks, S. Hu, Z. R. Kenz, et al., Model validation for a noninvasive arterial stenosis detection problem, CRSC-TR12-22, December, 2012; *Math. Biosci. Eng.*, **11** (2013), 427–448.
18. H. T. Banks and G. A. Pinter, A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue, CRSC-TR04-03, January, 2004; *SIAM J. Multiscale Model. Sim.*, **3** (2005), 395–412.
19. H. T. Banks, Chapter 14.4 of *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering,* Taylor and Frances Publishing, 2012.
20. G. de Vries, T. Hillen, M. Lewis, et al., *A Course in Mathematical Biology: Quantitative Modelling with Mathematical and Computational Methods*, SIAM, Philadephia, 2006.
21. S. I. Rubinow, *Introduction to Mathematical Biology*, John Wiley & Sons, New York, 1975.

22. L. J. Allen, *An Introduction to Mathematical Biology*. Pearson Education, Inc., Pearson Prentice Hall, Upper Saddle River, NJ, 2007.

23. M. Braun and M. Golubitsky, *Differential Equations and their Applications*. Vol. 4. Springer-Verlag New York, Inc., New York, NY, 1983.

24. G. E. Collins and A. G. Akritas, Polynomial real root isolation using Descarte's rule of signs, *Proceedings of the third ACM symposium on Symbolic and algebraic computation*, ACM, 1976.

25. H. T. Banks, L. Bekele-Maxwell, L. Bociu, et al., The complex-step method for sensitivity analysis of non-smooth problems arising in biology, CRSC-TR15-11, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, October, 2015; *Eurasia. J. Math. Comput. Appl.*, **3** (2015), 16–68.

26. H. T. Banks, K. Bekele-Maxwell, R. A. Everett, et al., Dynamic modeling of problem drinkers undergoing behavioral treatment, CRSC-TR16-12, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, October, October, 2016; *Bull. Math. Biol.*, **79** (2017), 1254–1273.

27. H. T. Banks and K. L. Bihari, Modeling and estimating uncertainty in parameter estimation, CRSC-TR99-40, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, October, 2016; December, 1999; *Inverse Probl.*, **17** (2001), 95–111.

28. H. T. Banks, K. B. Flores, I. G. Rosen, et al., The Prohorov Metric Framework and aggregate data inverse problems for random PDEs, CRSC-TR18-05, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, June, 2018; *Commun. Appl. Anal.*, **22** (2018), 415–446.

29. H. T. Banks, J. Catenacci and S. Hu, Use of difference-based methods to explore statistical and mathematical model discrepancy in inverse problems, CRSC-TR15-05, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, May, 2015. *J. Inverse Ill-pose. P.*, **24** (2016), 413-–433.

30. H. T. Banks, J. E. Banks, N. G. Cody, et al., Population model for the decline of *Homalodisca vitripennis* (HEMIPTERA: CICADELLIDAE) over a ten-year period, CRSC-TR18-06, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, June, 2018; *J. Biol. Dyn.*, **13** (2019), 422–446.

31. H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhasser, Boston, 1989.

32. H. T. Banks and P. Kareiva, Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models, *J. Math. Biol.*, **17** (1983), 253–273.

33. A. C. Atkinson and R. A. Bailey, One hundred years of the design of experiments on and off the pages of Biometrika, *Biometrika*, **88** (2001), 53–97.

34. A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*, Oxford University Press, New York, 1992.

35. M. P. F. Berger and W. K. Wong (Editors), *Applied Optimal Designs,* John Wiley & Sons, Chichester, UK, 2005.

36. V. V. Fedorov, *Theory of Optimal Experiments,* Academic Press, New York and London, 1972.

37. V. V. Fedorov and P. Hackel, *Model-Oriented Design of Experiments,* Springer-Verlag, New York, 1997.
38. W. Mueller and M. Stehlik, Issues in the optimal design of computer simulation experiments, *Appl. Stoch. Model. Bus.*, **25** (2009), 163–177.
39. M. Patan and B. Bogacka, Optimum experimental designs for dynamic systems in the presence of correlated errors, *Comput. Stat. Data An.*, **51** (2007), 5644–5661.
40. D. Ucinski and A. C. Atkinson, Experimental design for time-dependent models with correlated observations, *Stud. Nonlinear Dyn. E.*, **8** (2004), Article 13: The Berkeley Electronic Press.
41. H. T. Banks, K. Holm and F. Kappel, Comparison of optimal design methods in inverse problems, CRSC-TR10-11, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, July, 2010; *Inverse Probl.*, **27** (2011), 075002.
42. H. T. Banks, A. Cintr'on-Arias and F. Kappel, Parameter selection methods in inverse problem formulation, CRSC-TR10-03, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, revised November 2010; *Mathematical Model Development and Validation in Physiology: Application to the Cardiovascular and Respiratory Systems*, Lecture Notes in Mathematics, Vol. 2064, Mathematical Biosciences Subseries; Springer-Verlag, Berlin, 2013.
43. M. Avery, H. T. Banks, K. Basu, et al., Experimental design and inverse problems in plant biological modeling, CRSC-TR11-12, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, October, 2011; *J. Inverse Ill-pose. P.*, **20** (2012), 169–191.
44. H. T. Banks and K. L. Rehm, Experimental design for vector output systems, CRSC-TR12-11, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, April, 2012; *Inverse Probl. Sci. En.*, **22** (2014), 557–590.
45. H. T. Banks and K. L. Rehm, Experimental design for distributed parameter vector systems, CRSC-TR12-17, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, August, 2012; *Appl. Math. Lett.*, **26** (2013), 10–14.
46. H. T. Banks, S. Dediu, S. L. Ernstberger, et al., Generalized sensitivities and optimal experimental design, CRSC-TR08-12, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, September, 2008, revised November, 2009; *J. Inverse Ill-pose. P.*, **18** (2010), 25–83.
47. B. M. Adams, H. T. Banks, M. Davidian, et al., Model fitting and prediction with HIV treatment interruption data, CRSC TR05-40, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, October, 2005; *Bull. Math. Biol.*, **69** (2007), 563–584.

AIMS Press