ABSTRACT

KAPRAUN, DUSTIN FREDERICK. Cell Proliferation Models, CFSE-Based Flow Cytometry Data, and Quantification of Uncertainty. (Under the direction of H.T. Banks.)

The adaptive immune response is a major component of the human immune system's defense against invading pathogens. Since the success of the adaptive immune system depends on the capacity of lymphocytes to proliferate in response to various environmental stimuli (e.g., viral infection or organ transplant), the ability to make accurate predictions about lymphocyte behavior under specific conditions has important implications for immunology research. Such predictions can be made through the use of mathematical models.

In this dissertation, we provide an overview of our work using CFSE-based flow cytometry data in conjunction with cell proliferation models to estimate various biological quantities of interest. We first present a brief review of cell proliferation models that can be found in the literature and then describe two specific division- and label-structured PDE models which we have developed and used to analyze cell proliferation parameters. The first model we consider is based upon the premise of symmetric cell divisions, while the second allows for the possibility of asymmetric cell divisions. Considerable attention is devoted to derivations of the PDE models from conservation principles and explanations of the computational methods used to obtain numerical solutions for these models.

By examining a large collection of data sets involving replicated observations of CD4+ and CD8+ T cells collected from two healthy donors, we are able to analyze variability in CFSE-based flow cytometry data. Then, applying a statistical model and a generalized least squares parameter estimation scheme to our our symmetric division cell proliferation model, we are able to analyze variability in parameter estimates across multiple donors and cell types. We also discuss the identifiability of the various parameters involved in the symmetric division model. Finally, we apply a similar parameter estimation scheme to our asymmetric division model and observe and discuss differences in the results.

Cell Proliferation Models, CFSE-Based Flow Cytometry Data, and Quantification of Uncertainty

by
Dustin Frederick Kapraun

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2014

APPROVED BY:

Brian Reich	Ralph Smith
Hien Tran	H.T. Banks Chair of Advisory Committee

DEDICATION

To my sons, Henry Frederick Kapraun and Anders Joseph Kapraun, who give my life meaning; to my parents, who made this endeavor possible through their support, love, and encouragement; and to my wife, Jen, who inspired me to pursue a dream which I'd buried and almost forgotten.

É melhor ser alegre que ser triste
Alegria é a melhor coisa que existe
É assim como a luz no coração
Mas pra fazer um samba com beleza
É preciso um bocado de tristeza
Senão não se faz um samba não

– Vinicius de Moraes

BIOGRAPHY

The author was born in Wilmington, North Carolina, and spent his youth in Wilmington as well as in Raleigh, North Carolina, and the San Francisco Bay Area. He earned a B.S. in Mathematics (*summa cum laude*) in 1998 and an M.S. in Physics in 2002, both from N.C. State University. In between his undergraduate and master's degree studies, from 1998 to 1999, he worked as a computer programmer for a small company in Durham, North Carolina, and after attaining the master's degree he traveled extensively in South America. There he worked with spider monkeys at wildlife refuge in Bolivia, lived and worked at a small family-owned berry farm in Chile, and spent a month baking bread in the wilderness of central Brazil. Upon returning to the U.S., he served on the Long Trail Patrol in Vermont for several months before settling into a career in education. He taught math and statistics at Brunswick Community College in Supply, North Carolina, from 2004 to 2011 and served as Chair of the Math and Science Department at that school from 2007 to 2011. In June of 2011, he returned to N.C. State as a full-time student to pursue the Ph.D. in Applied Mathematics.

ACKNOWLEDGEMENTS

A number of individuals and entities have contributed to the work presented here, and without the knowledge, guidance, and financial support imparted by them, none of this would have been possible. In particular, I would like to thank my advisor, H.T. Banks, who, in addition to providing academic and professional direction, has been a constant source of encouragement, humor, and positivity. I am also deeply grateful to the members of my committee, Hien Tran, Ralph Smith, and Brian Reich, each of whom has contributed to my educational experience in more ways than one.

Now more than ever, applied math research depends on extensive multidisciplinary collaboration. It must be acknowledged that all experimental data used in this dissertation was supplied by our collaborators at Universitat Pompeu Fabra in Barcelona, Spain. These include Andreas Meyerhans, Jordi Argilaguet, and Cristina Peligero. It should also be mentioned that most of the computational results provided here rely heavily on MATLAB code authored by my collaborator and academic sibling W. Clay Thompson. I would also like to thank Katie Link, who participated in this project by helping to perform parameter estimations, and Jared Catenacci, who reviewed some of the mathematical proofs and arguments contained herein.

This research was supported in part by the National Institute of Allergy and Infectious Disease under grant NIAID R01AI071915-10 and in part by the National Science Foundation under Research Training Group (RTG) grant DMS-0636590 and Research Experience for Early Graduate Students (REG) grant DMS-0943855. I am also grateful for additional support in the form of a fellowship awarded by the Center for Research in Scientific Computation (CRSC) and Lord Corporation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

The adaptive immune response constitutes a major component of the mammalian immune system's defense against invading pathogens. B cells and T cells, which are the two classes of lymphocytes that comprise the adaptive immune system, recognize invaders by specific cell surface receptors and exert their responses through humoral and cellular effector mechanisms (i.e., through antibodies and cytotoxic T cells). As the success of the adaptive immune system in overcoming a perceived threat depends on the capacity of lymphocytes to proliferate, the ability to accurately predict cell proliferation dynamics in the presence of specific environmental stimuli has important implications for human health research in areas such as the treatment and prevention of infectious disease and immunosuppression for patients receiving organ and tissue transplants. Our efforts here focus on making such predictions through the use of mathematical models, and on quantifying uncertainty in these predictions.

Over the last half-century, many mathematical models have been proposed that attempt to describe the dynamics of a proliferating population of lymphocytes, but until fairly recently it has been a challenge to validate such models. The discovery of the intracellular dye carboxyfluorescein succinimidyl ester (CFSE) was a major milestone in overcoming this obstacle. CFSE was originally developed as a tool for labeling lymphocytes so that their movements within animal subjects could be tracked over many months [48], but subsequently researchers determined that the dye could also be used to monitor lymphocyte proliferation [36]. Also, through the use of fluorescently labeled antibodies specific to various lymphocyte surface markers, it is now possible to follow the proliferative behavior of specific types of lymphocytes [35]. We provide a detailed explanation of how CFSE and flow cytometry can be used to produce cell proliferation data in Section 1.2.

In the research presented here, CFSE-based flow cytometry data are used to determine T cell

proliferation and death rates (for cells having completed various numbers of divisions), and thus to obtain a more comprehensive understanding of the adaptive immune response. Some previous approaches to this problem are summarized in Section 1.3, and a then detailed description of our basic approach is presented in Chapter 2. It is worth noting that the mathematical and statistical methods described are sufficiently general that they could be applied to the analysis of various types of cells (not just lymphocytes) labeled with any intracellular fluorescent label (not just CFSE) that is apportioned approximately evenly upon mitosis. In Chapter 3, we describe a study in which variability in CFSE-based flow cytometry data, as well as experimental and biological variability in the associated proliferation parameters, are analyzed. In Chapter 4, we make a slight detour to analyze computational methods for computing convolution integrals associated with our cell proliferation models, and then in Chapter 5 we discuss a new model in which the apportioning of CFSE during mitosis is *not* assumed to be even.

1.2 CFSE-Based Flow Cytometry Data

The basic idea behind flow cytometry cell proliferation experiments can be described as follows. Suppose we can label all cells in a population with a "dye" that enters the cells and adheres to molecules within their cytoplasm. If we then stimulate the cells to divide, mitosis causes each "mother" cell to produce two "daughter" cells. Naturally, each of the daughter cells receives a portion of the cytoplasm that originally belonged to the mother, and therefore each of the daughter cells also receives a portion of the dye that was contained in the mother. Thus, as cell proliferation continues, those cells that have undergone more divisions tend to have less dye than those that have undergone fewer divisions. This allows one to use the total dye content of a cell as an indicator of the number of divisions it has undergone. In CFSE-based flow cytometry experiments, CFSE is the "dye" used to label the cells and flow cytometry is the means by which one can measure the total dye content of individual cells.

A variety of methods exist for preparing and culturing cells for use in flow cytometry experiments [35, 36, 39], but for the data utilized in this manuscript we consider the basic protocol illustrated in Figure 1.1. After collecting whole blood from subjects, the first step is to isolate the peripheral blood mononuclear cells (PBMCs) in the whole blood through centrifugation. PBMCs include lymphocytes, which in turn include the T cells upon which we focus our research efforts. Next, the PBMCs are passed through a filter that excludes clusters, or "clumps", of two or more cells. This proves important later, because the flow cytometer may not be able to distinguish a cluster of cells from a single cell containing an apparently large quantity of CFSE. The PBMCs, referred to hereafter as just "cells", are then exposed to CFSE. As described previously, the CFSE enters the cells and binds to proteins in the cytoplasm. After being "stained" in this way with CFSE, the cells are stimulated to divide. Phytohaemagglu-

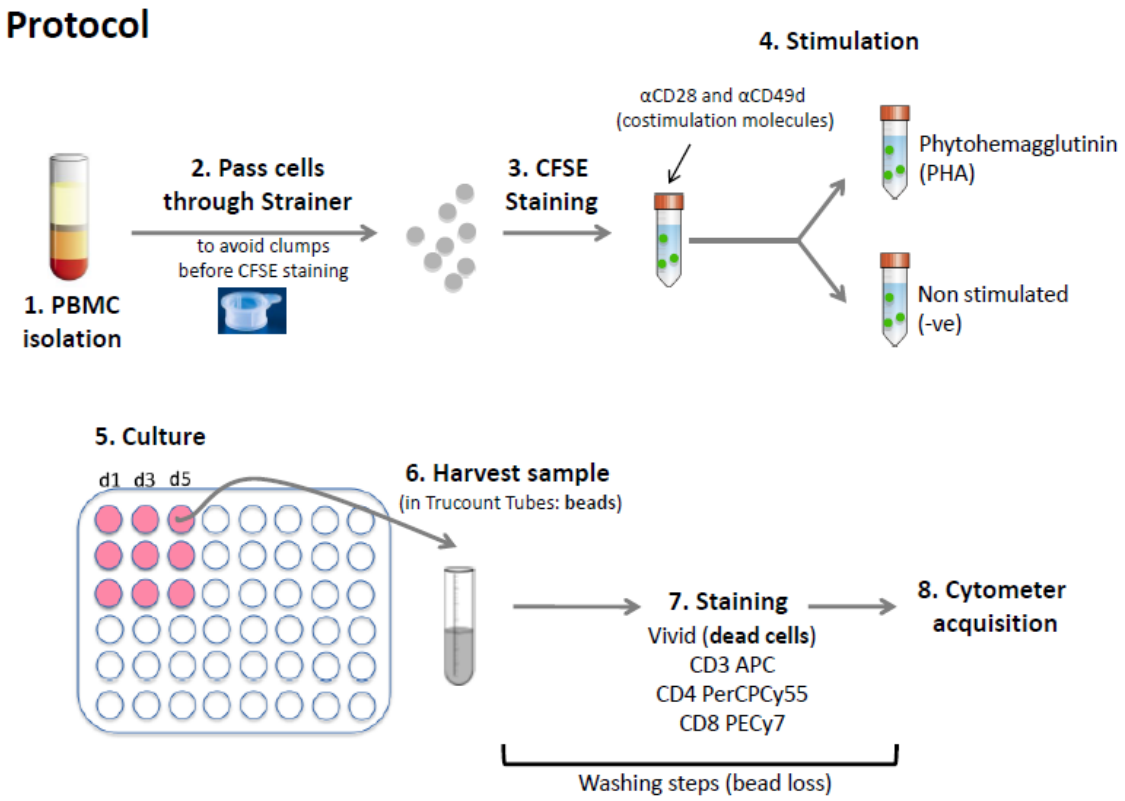


Figure 1.1: Cell preparation and culturing protocol. (Image courtesy of Cristina Peligero.)

tinin (PHA) is nonspecific T cell mitogen that can be used for this purpose. Figure 1.1 also shows a set of non-stimulated cells which can be used as an experimental control.

Once the cells have been thus prepared, they are placed into culturing wells with growth medium and left to proliferate. In Figure 1.1, we see that wells have been "seeded" in triplicate for each of three days. That is, three wells have been stocked with stained and stimulated cells in preparation for measurements that will occur on Day 1 ("d1"), and likewise for Day 3 ("d3") and Day 5 ("d5"). (The data sets considered in this manuscript actually consist of triplicate measurements made on each of *five* days.) It should be noted that experimenters take great care to seed each of the wells in an identical way. That is, each well is prepared with approximately the same number of cells and approximately the same quantity of growth media.

At the preordained time (e.g., after 24 hours, 72 hours, or 120 hours), the entire contents of a given well are placed into a sample tube that contains a known number of fluorescent beads. The cells in the tube are then exposed to other labeling agents which can be used to distinguish particular cell types of interest, such as "helper" (CD4+) T cells and cytotoxic or

"killer" (CD8+) T cells. Finally, a fraction (about 10 to 50%) of the contents of the sample tube is analyzed using flow cytometry. Before continuing with a brief description of flow cytometry, we think it worthwhile to list a few of the assumptions commonly made in conducting this type of cell proliferation experiment: (1) we assume that each well contains an identical population of cells at all times up until it is selected for measurement, (2) we assume that the sample of cells acquired in the flow cytometer is representative of the population of cells in the well from which the sample was taken, and (3) we assume that the fraction of the total well that is actually measured can be accurately estimated by comparing the number of beads that pass through the flow cytometer to the known total number of beads in the sample tube.

A flow cytometer is an instrument that is capable of quickly measuring various characteristics of large numbers of individual cells. As illustrated in Figure 1.2, this instrument uses hydrodynamic focusing to organize a sample of cells (and possibly other particles, such as beads) into a single file. The cells (and in our case also the beads) then pass one at a time through an interrogation point, where they are excited with a laser as shown in Figure 1.3. During this process, the flow cytometer measures the fluorescence intensity (FI) at various wavelengths for each cell in the sample and counts the number of beads in the sample. FI observed at wavelengths in the range 515 to 545 nm corresponds to light emitted by CFSE [35, 48], while FI observed at other wavelengths can indicate the presence of another labeling agent and can therefore be used to identify the type of cell. Because FI induced by CFSE varies directly with the mass of CFSE within a cell, the FI observed in the relevant range of wavelengths can be used as a surrogate for CFSE mass contained in a particular cell [35, 48].

The output of a CFSE-based flow cytometry experiment can be summarized using histograms such as those depicted in Figure 1.4. To construct these histograms, cells are placed into bins defined by ranges of CFSE FI. Because the measured FI numbers tend to vary over orders of magnitude during the course of typical multi-day experiment, it is common to use the base 10 logarithm of FI as in the figure. Note that each curve in the figure provides a summary of the information collected on a given day, and the abscissa and ordinate for each point on the curve correspond to the lower limit of a histogram bin and the number of cells belonging to that bin, respectively. Therefore, the total number of cells present in a well at a given point in time can be visualized as the total area under the curve corresponding to that time.

We hypothesize that each peak in one of the histograms represents a generation of cells having completed the same number of divisions. So in the "Day 1" histogram in Figure 1.4, we see a single peak representing a single generation of cells – presumably these are the undivided cells initially seeded in the well. In the "Day 2" histogram, however, we see two peaks because cell division has commenced. The peak on the right represents undivided cells, and the peak on the left represents cells that have divided once. Note that the FI corresponding to the center of the left peak on Day 2 (about $10^{4.2}$) is approximately one half of the FI corresponding the

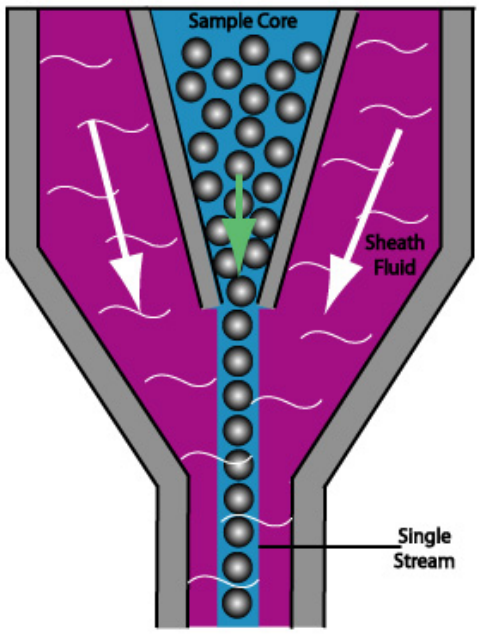


Figure 1.2: "Hydrodynamic focusing within the fluidics system of the flow cytometer." (Source: *www.selectscience.net*, 2 May 2014.)

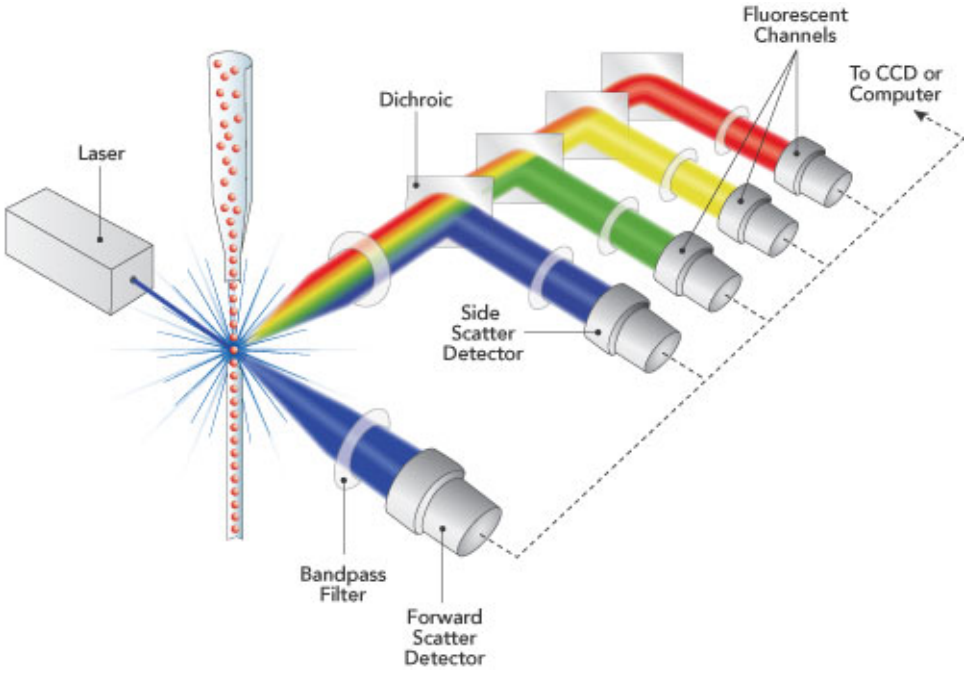


Figure 1.3: Overview of the flow cytometry apparatus. (Source: *www.semrock.com*, 2 May 2014.)

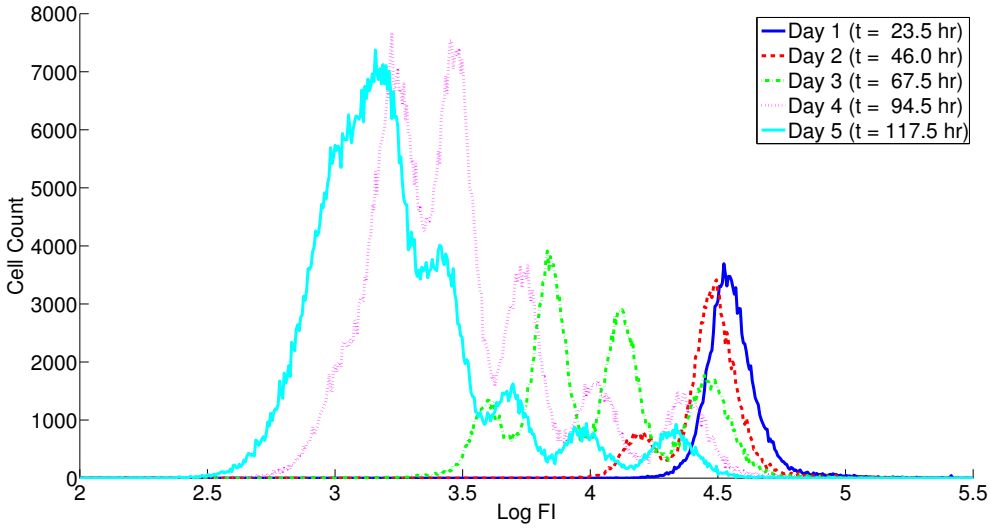


Figure 1.4: Histograms summarizing CFSE FI data at various time points.

center of the right peak (about $10^{4.5}$). This is to be expected if we assume that each daughter cell receives approximately one half of the CFSE contained in a dividing mother cell. Another point which is clearly demonstrated by Figure 1.4 is that the peak corresponding to a given generation tends to move to the left as time progresses. As described by Luzyanina et al. [34], this can be explained by a natural decay of the CFSE label that occurs within the cells. (See also references to CFSE "catabolism" by Lyons and Parish [36] and the discussion of CFSE "efflux" by Matera et al. [37].) Thus, there are two mechanisms by which the histograms tend to migrate to the left: cell division and CFSE label decay.

In the context of CFSE-based flow cytometry data, the goal of the modeling process is to link a mathematical description of cellular division and death processes at the population level to the observed fluorescence intensity profiles as measured by a flow cytometer (Figure 1.4). Because each peak in the flow cytometry data represents a cohort of cells having completed the same number of divisions, we hypothesize that flow cytometry data collected from cells stimulated to divide and then harvested at a series of time points will contain sufficient information to analyze the dynamic response of those cells to said stimulus. This dynamic response can only be accurately understood in the context of a mathematical model of the biological system that has been paired with a statistical model linking the mathematical model to the data. We hold that any such model must account for the decay of CFSE over time, the dilution of FI through cell division, and the asynchronous nature of the cellular division and death processes.

1.3 Earlier Cell Proliferation Models

Prior to the work of Luzyanina et al. in 2007 [34], most mathematical modeling of CFSE-based flow cytometry data focused on fitting numbers of cells per generation. To validate such models, researchers distill (or "deconvolute" [47]) summary histogram data such as that presented in Figure 1.4 to obtain the approximate numbers of cells in the various generations at each point in a time series. Deconvolution of label-based flow cytometry data is typically accomplished through interval gating or curve fitting techniques.

In interval gating, the FI (or log FI) axis of a summary histogram plot is partitioned into intervals which roughly correspond to peaks observed in the histogram [35, 38]. Recall that each peak corresponds to a generation of cells, so using this approach, the total number of cells observed in a particular interval at a particular point in time gives an estimate of the total number of cells in the corresponding generation at that time. Of course, such an approach presents problems when significant overlap occurs between peaks. This tends to happen when there is significant variation in the label content of cells in a given generation so that the peaks are relatively wide. In any case, overlap becomes more problematic with the later generations because after more divisions have occurred the total observable FI begins to approach the background FI (or "autofluorescence", cf. Section 2.2) of the cells in question. That is, once cells have divided a certain number of times, say 10, autofluorescence dominates their observable FI. Thus, the FI distributions for cells that have divided 10 times versus 12 times are usually indistinguishable.

Curve fitting forms the basis for a more sophisticated approach to deconvolution of label-based flow cytometry data. In this technique, one attempts to find a series of Gaussian or lognormal curves that yield a good approximation of the summary histogram data [26, 35, 39]. Computer algorithms designed for this purpose usually employ a least squares approach, and a number of commercially available software packages with this functionality now exist. Figure 1.5 shows typical output for one such software package. In blue, we see a set of 8 scaled lognormal probability density functions. Each of these curves corresponds to a cohort of cells in the same generation, so labels are provided across the top of the figure indicating cells that have divided "0" times, "1 time", and so on. In red, we see the curve given by the sum of the 8 blue curves. It is this curve that presumably gives the best fit to the non-smooth black curve, which itself represents the summary histogram data. For the example depicted, the total number of cells in each of 8 generations can be approximated by determining the area under the corresponding blue lognormal curve.

A variety of models have been proposed for describing the numbers of cells per generation at a given point in time, and most of these can be described by systems algebraic equations, ordinary differential equations (ODEs), or combinations thereof [26, 22, 23, 38]. These models

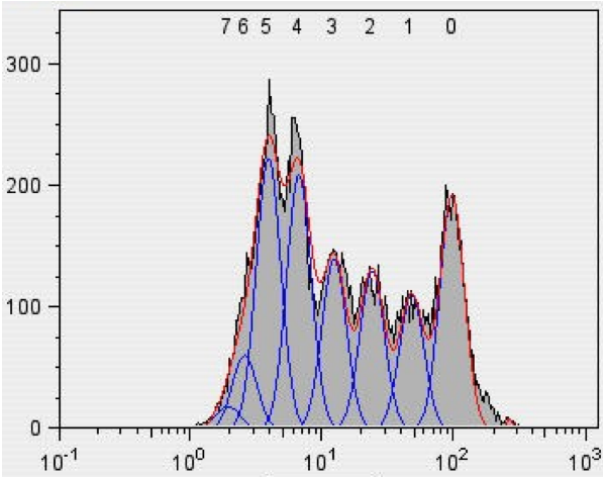


Figure 1.5: Deconvolution of summary histogram data. (Source: *www.flojo.com*, 27 June 2014.)

establish biologically meaningful parameters such as division and death rates and time to first division, and show how they can be identified using label-based flow cytometry data sets. More recently, Hawkins et al. introduced a novel approach to modeling numbers of cells per generation in which the parameters to be estimated are actually distributions [28, 29]. This model, which they call the "cyton" model, is constructed upon the hypothesis that the fate of any given cell is determined based upon realizations of two random variables: a time to divide and a time to die. We give more attention to the cyton model for cell numbers in Section 2.3.

Each of the models cited thus far has been used to provide measures of the proliferative capacity of a population of cells, but we again emphasize that these models are generally calibrated using deconvoluted data. Clearly, the assumption of particular distributional shapes for the generational structure of FI histogram data (e.g., the assumption that the generational peaks have lognormal distributions) affects how cell numbers per generation are estimated, and can therefore introduce bias into the parameter estimates obtained in model calibrations. Furthermore, because the original "raw" data are essentially discarded once the cell number estimates are computed, any attempt to develop a statistical model linking a mathematical model that only describes cell numbers to the data has been marred from the outset.

In recent years, a new type of structured cell population model has emerged which promises to overcome some of the issues we've raised concerning cell number models and deconvoluted data. To our knowledge, Luzyanina et al. [34] proposed the first partial differential equation (PDE) model to employ fluorescence intensity (FI) as a structure variable. They showed that such a model can successfully track the behavior of a proliferating lymphocyte population labeled with CFSE, and that it can compete with compartmental ODE models in estimating cell

numbers per generation. More recent work with FI- or label-structured PDE models [1, 9, 33] has consistently demonstrated their effectiveness in fitting summary FI histogram data for CFSE-based proliferation assays. The key idea behind such models is that, because the CFSE-based FI tends to be less for cells that have divided more times, FI can be used as a surrogate for division number. By directly treating measured FI data and the experimental processes underlying flow cytometry and CFSE-based proliferation assays, structured PDE models offer clear advantages over earlier cell number models.

The model originally put forth by Luzyanina et al. describes the number of cells per unit FI at a given time in terms of a single hyperbolic PDE. Unfortunately, with this approach one must still associate a particular interval of FI with a specific generation of cells in order to incorporate generational dependence into division and death rate functions. Thus, the problem of overlap between generational peaks persists. More recently, Thompson [47] proposed a compartmental model consisting of one PDE for each generation under consideration. This model eliminates the need to define each generation of cells through specification of an FI interval, and is the starting point for the cell proliferation models we discuss in Chapter 2.

Chapter 2

Modeling Cell Proliferation with Partial Differential Equations

Here, we summarize a class of models originally proposed by Thompson [47] and further developed by Banks, Thompson, et al. [10, 5]. We then identify a specific mathematical model for further consideration, describe a statistical model relating this mathematical model to CFSE-based flow cytometry data, and outline a scheme by which model parameters can be estimated.

2.1 Basic Division- and Label-Structured Model for Cell Densities

Let $n_i(t,x)$ be a structured density (in cells per unit FI), where i is a whole number representing the number of divisions completed for a specific "generation" of cells, t denotes the time elapsed (in hr) since some arbitrary starting time, and x denotes FI *induced by CFSE*. Also, let $\{\alpha_i(t)\}$, $\{\beta_i(t)\}$, and $v(t)$ denote exponential division rates, exponential death rates, and the CFSE exponential decay rate, respectively (all in hr^{-1}). Then the dynamics of a population of cells are described by

$$\begin{aligned} \frac{\partial}{\partial t}n_0(t,x) - v(t)\frac{\partial}{\partial x}\Big[xn_0(t,x)\Big] &= -\Big(\alpha_0(t)+\beta_0(t)\Big)n_0(t,x) && (\text{for } i=0), \\ \frac{\partial}{\partial t}n_i(t,x) - v(t)\frac{\partial}{\partial x}\Big[xn_i(t,x)\Big] &= -\Big(\alpha_i(t)+\beta_i(t)\Big)n_i(t,x) + R_i(t,x) && \text{for } i\geq 1, \end{aligned} \tag{2.1}$$

where $x \geq 0$ and the "recruitment" terms are given by

$$R_i(t,x) = 4\alpha_{i-1}(t)n_{i-1}(t,2x) \tag{2.2}$$

for $i \geq 1$. The initial conditions are given by

$$n_i(t_0, x) = \begin{cases} \Phi(x) & \text{for } i = 0, \\ 0 & \text{for } i \geq 1, \end{cases} \tag{2.3}$$

where t_0 indicates the time of the first observation and $\Phi(x)$ is the structured density for cells in the initial (undivided) population. We remark here that t_0 typically coincides with the time at which the cells were stimulated to divide, but for the experimental data we describe in Chapter 3 and consider throughout this dissertation, the first observation actually occurred approximately 24 hours after stimulation. We also mention that the form of the recruitment terms (2.2) assumes an even partitioning of the CFSE in a mother cell between two daughter cells during cytokinesis; i.e., *we assume that each daughter cell receives exactly one half of the CFSE that was present in the mother cell.* It should be pointed out that long-standing results indicate that the partitioning of cytoplasm to two daughter cells during mitosis is *not* even [44], and a recent review [15] suggests that incorporating the assumption of asymmetric cell division into mathematical models for cell proliferation will "improve assessment of T cell performance parameters from CFSE-based proliferation assays." In Chapter 5 we revisit the possibility of asymmetric division, but here we follow the convention of earlier work [5, 27, 35, 39, 41, 47] in making the simplifying assumption that CFSE is evenly distributed during cell division.

The derivation of a model very similar to that given in (2.1) and (2.2) was provided in Chapter 3 of [47], but there the contribution of cellular autofluorescence was treated differently than it will be treated here (see Section 2.2). For that reason, and for the sake of completeness, we offer our own model derivation presently. In order to obtain a partial differential equation describing $n_i(t, x)$, we consider changes to the total number of cells in generation i at time t with CFSE FI in the arbitrary interval $[x, x + \Delta x]$. This total number of cells is given by

$$\int_x^{x+\Delta x} n_i(t, \xi)\, d\xi.$$

There are five possible contributions to the time rate of change of this quantity, and we consider each of these in the list below.

(i) The rate at which cells enter the FI interval $[x, x + \Delta x]$ (from the right) due to CFSE decay, which can be computed as $v(t) \cdot (x + \Delta x) \cdot n_i(t, x + \Delta x)$.

(ii) The rate at which cells leave the FI interval $[x, x + \Delta x]$ (from the left) due to CFSE decay, which can be computed as $v(t) \cdot x \cdot n_i(t, x)$.

(iii) The rate at which cells leave the FI interval $[x, x + \Delta x]$ (and, in fact, leave generation i)

due to cell division, which can be computed as

$$\alpha_i(t) \cdot \int_x^{x+\Delta x} n_i(t,\xi)\, d\xi.$$

(iv) The rate at which cells leave the FI interval $[x, x+\Delta x]$ due to cell death, which can be computed as

$$\beta_i(t) \cdot \int_x^{x+\Delta x} n_i(t,\xi)\, d\xi.$$

(v) The rate at which cells enter the FI interval $[x, x+\Delta x]$ due to the division of cells in the "previous" generation $i-1$ with FI in the interval $[2x, 2(x+\Delta x)]$. As mentioned above, we assume that each daughter cell receives exactly one half of the CFSE present in the mother cell during mitosis. Because two daughter cells are created from a single mother cell, the rate in question can be computed as 2 times the exponential birth rate times the number of cells in the relevant FI interval, or

$$2 \cdot \alpha_{i-1}(t) \cdot \int_{2x}^{2(x+\Delta x)} n_{i-1}(t,\xi)\, d\xi.$$

Applying the change of variables $\eta = \frac{1}{2}\xi$, the expression becomes

$$4 \cdot \alpha_{i-1}(t) \cdot \int_x^{x+\Delta x} n_{i-1}(t,2\eta)\, d\eta.$$

It is important to note that this particular contribution to change in cell numbers *does not apply* in the case of cells in generation $i = 0$ because there is no previous generation from which cells can enter in that case.

Taking into account all of these contributions, the time rate of change of the total number of cells in generation i at time t with FI in the region $[x, x+\Delta x]$ is

$$\begin{aligned}
\frac{d}{dt}\int_x^{x+\Delta x} n_i(t,\xi)\, d\xi &= \text{(i)} - \text{(ii)} - \text{(iii)} - \text{(iv)} + \text{(v)} \\
&= \Big[v(t)(x+\Delta x) n_i(t, x+\Delta x)\Big] - \Big[v(t) x n_i(t,x)\Big] \\
&\quad - \left[\alpha_i(t) \cdot \int_x^{x+\Delta x} n_i(t,\xi)\, d\xi\right] - \left[\beta_i(t) \cdot \int_x^{x+\Delta x} n_i(t,\xi)\, d\xi\right] \\
&\quad + 4\alpha_{i-1}(t) \int_x^{x+\Delta x} n_{i-1}(t,2\eta)\, d\eta.
\end{aligned}$$

Dividing by Δx on both sides of this equation and taking the limit as $\Delta x \to 0$ yields

$$\frac{\partial}{\partial t} n_i(t,x) = v(t)\frac{\partial}{\partial x}\Big[x n_i(t,x)\Big] - \Big(\alpha_i(t) + \beta_i(t)\Big) n_i(t,x) + 4\alpha_{i-1}(t) n_{i-1}(t,2x).$$

As mentioned above, the last term on the right hand side of this equation must be omitted when $i = 0$. Thus, we obtain the model given in (2.1) and (2.2).

As proposed by Schittler et al. [41], the solutions to the partial differential equations (PDEs) given in (2.1) can be factored as

$$n_i(t,x) = N_i(t)\bar{n}_i(t,x), \tag{2.4}$$

where $N_i(t)$ indicates the *number* of cells having completed i divisions at time t and $\bar{n}_i(t,x)$ describes the *distribution* of CFSE FI within that generation of cells at time t; that is, $\bar{n}_i(t,x)$ is a probability density function (pdf) in the variable x, so that for any fixed t, $\bar{n}_i(t,x) \geq 0$ for all x and

$$\int_0^\infty \bar{n}_i(t,x)\, dx = 1.$$

Because the assumptions and notations we employ here are slightly different from those used by Schittler et al. [41], we offer the following formal proposition of this factorability of solutions and provide a proof.

Proposition 2.1. *Let $\{N_i(t)\}_{i=0}^\infty$ be a set of functions satisfying the system of weakly coupled ordinary differential equations (ODEs) given by*

$$\begin{aligned} \frac{dN_0(t)}{dt} &= -\Big(\alpha_0(t) + \beta_0(t)\Big) N_0(t) && (\text{for } i = 0), \\ \frac{dN_i(t)}{dt} &= -\Big(\alpha_i(t) + \beta_i(t)\Big) N_i(t) + 2\alpha_{i-1}(t) N_{i-1}(t) && \text{for } i \geq 1, \end{aligned} \tag{2.5}$$

and the initial conditions given by

$$N_i(t_0) = \begin{cases} N_0 = \int_0^\infty \Phi(x)\, dx & \text{for } i = 0, \\ 0 & \text{for } i \geq 1. \end{cases} \tag{2.6}$$

Also, let $\{\bar{n}_i(t,x)\}_{i=0}^\infty$ be a set of functions such that each $\bar{n}_i$ satisfies the PDE

$$\frac{\partial \bar{n}_i(t,x)}{\partial t} - v(t)\frac{\partial [x\bar{n}_i(t,x)]}{\partial x} = 0 \tag{2.7}$$

and the initial condition

$$\bar{n}_i(t_0,x) = \frac{2^i \Phi(2^i x)}{N_0} \tag{2.8}$$

for all $x \geq 0$. Then the solution to (2.1) *and* (2.3) *is given by* (2.4) *for $i \in \{0, 1, 2, \ldots\}$.*

Proof. We begin by considering the case $i = 0$. Taking the time derivative of (2.4) and then substituting (2.5) and (2.7) leads to

$$\begin{aligned}
\frac{\partial}{\partial t} n_0(t,x) &= \frac{d}{dt}\Big[N_0(t)\Big] \cdot \bar{n}_0(t,x) + N_0(t) \cdot \frac{\partial}{\partial t}\Big[\bar{n}_0(t,x)\Big] \\
&= \Big[-\Big(\alpha_0(t) + \beta_0(t)\Big) N_0(t)\Big] \cdot \bar{n}_0(t,x) + N_0(t) \cdot \left[v(t) \frac{\partial}{\partial x}\Big[x \bar{n}_0(t,x)\Big]\right] \\
&= -\Big(\alpha_0(t) + \beta_0(t)\Big) N_0(t) \bar{n}_0(t,x) + v(t) \frac{\partial}{\partial x}\Big[x N_0(t) \bar{n}_0(t,x)\Big] \\
&= -\Big(\alpha_0(t) + \beta_0(t)\Big) n_0(t,x) + v(t) \frac{\partial}{\partial x}\Big[x n_0(t,x)\Big],
\end{aligned}$$

which is equivalent to (2.1) in the case $i = 0$. Furthermore, substituting (2.6) and (2.8) into (2.4), the initial condition for the case $i = 0$ becomes

$$n_0(t_0, x) = N_0(t_0) \bar{n}_0(t_0, x) = \big[N_0\big] \cdot \left[\frac{\Phi(x)}{N_0}\right] = \Phi(x),$$

which is equivalent to (2.3) in the case $i = 0$.

Next, we consider the situation for $i \geq 1$. For this case, it is first necessary to obtain the solutions of (2.7) subject to the initial conditions (2.8). These solutions, which can be obtained by the method of characteristics (see Section A.1 of the Appendix), are given by

$$\bar{n}_i(t,x) = \frac{2^i}{N_0} \Phi\left(2^i x \cdot \exp\left[\int_{t_0}^{t} v(u)\, du\right]\right) \cdot \exp\left[\int_{t_0}^{t} v(u)\, du\right].$$

From the expressions for these solutions it is easily verified (see Lemma A.1) that

$$\bar{n}_i(t,x) = 2 \bar{n}_{i-1}(t, 2x) \tag{2.9}$$

for $i \geq 1$, a result which we shall use presently.

For an index $i \geq 1$, taking the time derivative of (2.4) and then substituting (2.5) and (2.7) leads to

$$\begin{aligned}
\frac{\partial}{\partial t} n_i(t,x) &= \frac{d}{dt}\Big[N_i(t)\Big] \cdot \bar{n}_i(t,x) + N_i(t) \cdot \frac{\partial}{\partial t}\Big[\bar{n}_i(t,x)\Big] \\
&= \Big[-\Big(\alpha_i(t) + \beta_i(t)\Big) N_i(t) + 2\alpha_{i-1}(t) N_{i-1}(t)\Big] \cdot \bar{n}_i(t,x) + N_i(t) \cdot \left[v(t) \frac{\partial}{\partial x}\Big[x \bar{n}_i(t,x)\Big]\right] \\
&= -\Big(\alpha_i(t) + \beta_i(t)\Big) N_i(t) \bar{n}_i(t,x) + 2\alpha_{i-1}(t) N_{i-1}(t) \bar{n}_i(t,x) + v(t) \frac{\partial}{\partial x}\Big[x N_i(t) \bar{n}_i(t,x)\Big].
\end{aligned}$$

Then, making the substitution suggested by (2.9) in the second term on the right hand side,

we obtain

$$\begin{aligned}\frac{\partial}{\partial t}n_i(t,x) &= -\Big(\alpha_i(t)+\beta_i(t)\Big)N_i(t)\bar{n}_i(t,x) + 4\alpha_{i-1}(t)N_{i-1}(t)\bar{n}_{i-1}(t,2x) + v(t)\frac{\partial}{\partial x}\Big[xN_i(t)\bar{n}_i(t,x)\Big] \\ &= -\Big(\alpha_i(t)+\beta_i(t)\Big)n_i(t,x) + 4\alpha_{i-1}(t)n_{i-1}(t,2x) + v(t)\frac{\partial}{\partial x}\Big[xn_i(t,x)\Big]\end{aligned}$$

which is equivalent to (2.1) in the case $i \geq 1$. Substituting (2.6) and (2.8) into (2.4), the initial condition for an index $i \geq 1$ becomes

$$n_i(t_0,x) = N_i(t_0)\bar{n}_i(t_0,x) = \big[0\big]\cdot\left[\frac{2^i\Phi(2^i x)}{N_0}\right] = 0,$$

which is equivalent to (2.3) in the case $i \geq 1$.

Thus, we have verified that the solution to (2.1) and (2.3) is given by (2.4) for $i \in \{0,1,2,\ldots\}$, provided the conditions stipulated in the proposition are met. □

It is worth noting that cells will only divide a finite number of times in the time frame of a typical *in vitro* cell culturing experiment. Therefore, we typically compute solutions to (2.5) and (2.7) only for $i \in \{0,1,\ldots,i_{\max}\}$, where $i_{\max}$ is the largest number of divisions we expect a cell from the initial population to undergo during the period of observation. For a five-day experiment such as that presented in Section 3.1, it is rare for cells to undergo more than 12 divisions. In order to capture the behavior of all but a negligible number of cells, we therefore use the conservative value of $i_{\max} = 16$ for our purposes in Chapter 3 and elsewhere in this dissertation unless another value is explicitly noted.

2.2 Autofluorescence

Thus far, we have described a model that accounts only for FI *induced by CFSE*, but as noted in [47], the experimentally measured FI of a cell is actually the sum of CFSE-induced FI and the cell's natural "autofluorescence". Therefore, following the work of [27], we let $\tilde{n}_i(t,\tilde{x})$ be a structured density (in cells per unit FI), where i again denotes a specific generation of cells, t denotes time elapsed (in hr), and $\tilde{x}$ denotes *measured* FI. Here,

$$\tilde{x} = x + x_a,$$

where x and x_a represent the FI due to CFSE content and cellular autofluorescence, respectively.

If we assume solutions $n_i(t,x)$ to (2.1) and (2.3) have already been computed and that x_a is a realization of a random variable X_a with pdf $f_{X_a}(x_a;t)$, then the densities $\tilde{n}_i(t,\tilde{x})$ can be

computed using the convolution formula [18]

$$\tilde{n}_i(t,\tilde{x}) = \int_{-\infty}^{\infty} n_i(t,x) f_{X_a}(\tilde{x}-x;t)\,dx = \int_0^{\tilde{x}} n_i(t,x) f_{X_a}(\tilde{x}-x;t)\,dx. \tag{2.10}$$

as proposed by Hasenauer, Schittler, et al. [27, 41]. These authors demonstrate that, under certain assumptions, this convolution can be computed quickly and efficiently [27]. More will be said about the computation of (2.10) in Chapter 4.

2.3 Cyton Model for Cell Numbers

We now turn our attention to the cyton model [28, 29], which is an alternative to (2.5) that arises from two simple assumptions. The first, which is self-evident, is that any given cell must eventually either divide or die. The second, which is based upon experimental evidence, is that the processes of cell division and death operate independently of one another [29]. Thus, we can assume that the destiny of any particular cell is governed by two fixed numbers: a "time until division" and a "time until death". In particular, the actual fate of the cell (division or death) can be determined by observing which of these two numbers is smaller. For an individual cell within a population of cells sharing similar characteristics (e.g., cells of the same type having undergone the same number of divisions), it is reasonable to assume that the "time until division" and "time until death" are realizations of independent random variables. These random variables are described by probability distributions, and so the cyton model requires parameters that can be used to uniquely determine the probability distributions for times until division and death of cells in a given population (e.g., CD4+ T cells having undergone 1 division). Hawkins et al. chose the term "cyton" to denote the "combination of independent cellular machines governing times to divide and die" and represented a cyton mathematically using a pair of probability density functions [29] . For example, if ϕ_i and ψ_i are the pdfs for time until division and time until death, respectively, of cells in generation i, then the cyton for generation i can be denoted (ϕ_i, ψ_i). One additional consideration is that, in reality, not all cells in a given population will divide if they avoid death (at least not within the time frame of a typical *in vitro* cell culturing experiment) [29]. Therefore, the cyton model includes the notion of "progressor fraction": for a given generation of cells, only a certain proportion have the potential to "progress" to the next generation via cell division.

Let F_i denote the progressor fraction for generation i; that is, F_i represents the proportion of cells in generation i that would (eventually) divide in the absence of any possibility of cell death. Then, define the random variable T_i^{div} to be the time required for a cell in generation i (with the *potential* to progress) to complete its next division (measured in hours since the completion of the i^{th} division, or in the case of T_0^{div}, hours since t_0). Similarly, define the random

variable T_i^{die} to be the time required for a cell in generation i to die. Finally, let $\phi_i(t)$ and $\psi_i(t)$ be pdfs for T_i^{div} and T_i^{die}, respectively. If we define $N_i(t)$ as before, the cyton model is then given by

$$\begin{aligned} N_0(t) &= N_0 - \int_{t_0}^{t} \left(n_0^{div}(s) + n_0^{die}(s) \right) ds \qquad (\text{for } i = 0), \\ N_i(t) &= \int_{t_0}^{t} \left(2n_{i-1}^{div}(s) - n_i^{div}(s) - n_i^{die}(s) \right) ds \quad \text{for } i \geq 1, \end{aligned} \tag{2.11}$$

where $n_i^{div}(t)$ and $n_i^{die}(t)$ are rates (in cells/hr) at which cells in generation i divide and die, respectively. These rates are defined as

$$n_i^{div}(t) = \begin{cases} F_0 N_0 \left(1 - \displaystyle\int_{t_0}^{t} \psi_0(s)\, ds \right) \phi_0(t) & \text{for } i = 0, \\ 2F_i \displaystyle\int_{t_0}^{t} n_{i-1}^{div}(s) \left(1 - \int_0^{t-s} \psi_i(\xi)\, d\xi \right) \phi_i(t-s)\, ds & \text{for } i \geq 1. \end{cases} \tag{2.12}$$

and

$$n_i^{die}(t) = \begin{cases} N_0 \left(1 - F_0 \displaystyle\int_{t_0}^{t} \phi_0(s) ds \right) \psi_0(t) & \text{for } i = 0, \\ 2 \displaystyle\int_{t_0}^{t} n_{i-1}^{div}(s) \left(1 - F_i \int_0^{t-s} \phi_i(\xi)\, d\xi \right) \psi_i(t-s)\, ds & \text{for } i \geq 1. \end{cases} \tag{2.13}$$

There is considerable experimental evidence [11, 28, 29] that supports the cyton model, and it has an advantage over models such as (2.5) in that it directly connects cell population numbers to probablity distributions describing times at which cells in a given generation will divide or die. Identifying these distributions for populations of lymphocytes exposed to specific environmental stimuli allows for a detailed quantitative description of the adaptive immune response.

2.4 Division- and Label-Structured Cyton Model for Cell Densities

As in earlier work [5], we incorporate the cyton model for cell numbers into the division- and label-structured model framework described previously by replacing the sink and source terms in the right-hand sides of (2.1) with terms involving the cyton-based rates to obtain

$$\begin{aligned} \frac{\partial n_0(t,x)}{\partial t} - v(t) \frac{\partial [x n_0(t,x)]}{\partial x} &= -\Big(n_0^{div}(t) + n_0^{die}(t) \Big) \bar{n}_0(t,x) \qquad (\text{for } i = 0), \\ \frac{\partial n_i(t,x)}{\partial t} - v(t) \frac{\partial [x n_i(t,x)]}{\partial x} &= \Big(2n_{i-1}^{div}(t) - n_i^{div}(t) - n_i^{die}(t) \Big) \bar{n}_i(t,x) \quad \text{for } i \geq 1. \end{aligned} \tag{2.14}$$

Solutions of this system are then given by $n_i(t,x) = N_i(t)\bar{n}_i(t,x)$, where the $N_i(t)$'s satisfy the cyton model equations (2.11) and initial conditions (2.6) and each $\bar{n}_i(t,x)$ satisfies (2.7) and (2.8) as before. We state this claim as a proposition and provide a proof below.

Proposition 2.2. *Let $\{N_i(t)\}_{i=0}^{\infty}$ be a set of functions satisfying the cyton model (2.11), where the initial condition N_0 is given by the relation shown in (2.6). Also, let $\{\bar{n}_i(t,x)\}_{i=0}^{\infty}$ be a set of functions such that each $\bar{n}_i$ satisfies the PDE (2.7) and the initial condition (2.8) for all $x \geq 0$. Then the solution to (2.14) with initial conditions (2.3) is given by (2.4) for $i \in \{0,1,2,\ldots\}$.*

Proof. We begin by considering the case in which $i = 0$. Taking the time derivative of (2.4), substituting (2.11) and (2.7), and applying the fundamental theorem of calculus leads to

$$\begin{aligned}
\frac{\partial}{\partial t}n_0(t,x) &= \frac{d}{dt}\Big[N_0(t)\Big]\cdot\bar{n}_0(t,x) + N_0(t)\cdot\frac{\partial}{\partial t}\Big[\bar{n}_0(t,x)\Big] \\
&= \frac{d}{dt}\left[N_0 - \int_{t_0}^{t}\Big(n_0^{div}(s) + n_0^{die}(s)\Big)\,ds\right]\cdot\bar{n}_0(t,x) + N_0(t)\cdot\left[v(t)\frac{\partial}{\partial x}\Big[x\bar{n}_0(t,x)\Big]\right] \\
&= \Big[0 - \Big(n_0^{div}(t) + n_0^{die}(t)\Big)\Big]\cdot\bar{n}_0(t,x) + v(t)\frac{\partial}{\partial x}\Big[xN_0(t)\bar{n}_0(t,x)\Big] \\
&= -\Big(n_0^{div}(t) + n_0^{die}(t)\Big)\bar{n}_0(t,x) + v(t)\frac{\partial}{\partial x}\Big[xn_0(t,x)\Big],
\end{aligned}$$

which is equivalent to (2.14) in the case $i = 0$. Furthermore, evaluating (2.4) at $t = t_0$ and substituting (2.11) and (2.8), the initial condition for the case $i = 0$ becomes

$$n_0(t_0,x) = N_0(t_0)\bar{n}_0(t_0,x) = \big[N_0\big]\cdot\left[\frac{\Phi(x)}{N_0}\right] = \Phi(x),$$

which is equivalent to (2.3) in the case $i = 0$.

Next, we consider the situation for $i \geq 1$. As for the $i = 0$ case, we take the time derivative of (2.4), substite (2.11) and (2.7), and apply the fundamental theorem of calculus. The result in this case is

$$\begin{aligned}
\frac{\partial}{\partial t}n_i(t,x) &= \frac{d}{dt}\Big[N_i(t)\Big]\cdot\bar{n}_i(t,x) + N_i(t)\cdot\frac{\partial}{\partial t}\Big[\bar{n}_i(t,x)\Big] \\
&= \frac{d}{dt}\left[\int_{t_0}^{t}\Big(2n_{i-1}^{div}(s) - n_i^{div}(s) - n_i^{die}(s)\Big)\,ds\right]\cdot\bar{n}_i(t,x) + N_i(t)\cdot\left[v(t)\frac{\partial}{\partial x}\Big[x\bar{n}_i(t,x)\Big]\right] \\
&= \Big(2n_{i-1}^{div}(t) - n_i^{div}(t) - n_i^{die}(t)\Big)\cdot\bar{n}_i(t,x) + v(t)\frac{\partial}{\partial x}\Big[xN_i(t)\bar{n}_i(t,x)\Big] \\
&= \Big(2n_{i-1}^{div}(t) - n_i^{div}(t) - n_i^{die}(t)\Big)\bar{n}_i(t,x) + v(t)\frac{\partial}{\partial x}\Big[xn_i(t,x)\Big],
\end{aligned}$$

which is equivalent to (2.14) in the case $i \geq 1$. Finally, evaluating (2.4) at $t = t_0$ and substituting

(2.11) and (2.8), the initial condition for the case $i \geq 1$ becomes

$$n_i(t_0, x) = N_i(t_0)\bar{n}_i(t_0, x) = \big[0\big] \cdot \left[\frac{2^i \Phi(2^i x)}{N_0}\right] = 0,$$

which is equivalent to (2.3) in the case $i \geq 1$.

Thus, we have verified that the solution to (2.14) and (2.3) is given by (2.4) for $i \in \{0, 1, 2, \ldots\}$, provided the conditions stipulated in the proposition are met. □

Like (2.1), the model given by (2.14) may be properly described as a division- and label-structured population model, as it makes use of structure variables for division number (or generation) i and CFSE-induced FI x (which is assumed to be proportional to CFSE label content). Also, as described by Hasenauer, Schittler, et al. [27, 41] and summarized in our earlier work [5], the factorable form of the solutions $\{n_i(t, x)\}$ and the technique for converting these to corresponding solutions $\{\tilde{n}_i(t, \tilde{x})\}$ via convolution integrals (cf. (2.10)) makes it possible to obtain numerical solutions very quickly when the model parameters are given. This model has been shown to yield a reasonably good fit to summary histogram data such as that presented in Figure 1.4, provided that the model parameters are chosen "optimally" [5]. More will be said about optimal parameter estimation in Section 2.6.

Finally, we remark that (2.14) actually describes an entire class of models. In order to specify a particular model for further investigation, we must decide on forms for the distribution of the autofluorescence X_a, the (exponential) label decay rate $v(t)$, the cytons $\{(\phi_i(t), \psi_i(t))\}$, and the progressor fractions $\{F_i\}$.

2.5 Assumptions and Parameterization for a Specific Mathematical Model

Here, we list the assumptions for the specific cyton-based mathematical model we consider and describe the parameters that we use to designate this model. All of the parameters for our specific mathematical model are provided in Table 2.1.

First, we assume that the random variable X_a is time-independent and has a lognormal distribution with mean and variance denoted $\mathrm{E}[X_a]$ and $\mathrm{Var}[X_a]$, respectively. Although experiments indicate that the distribution of autofluorescence does, in fact, vary with time, ignoring this time-dependence greatly reduces the number of parameters required to designate the model and still allows for a reasonable fit to summary histogram data [5]. We therefore have two parameters that completely describe the distribution of autofluorescence: $\mathrm{E}[X_a]$ and $\mathrm{SD}[X_a] = (\mathrm{Var}[X_a])^{1/2}$, which are listed as parameters 1 and 2, respectively, in Table 2.1.

Next, we assume that the rate of decay in CFSE-induced FI is given by $v(t) = c$, where $c > 0$ is some constant. This follows the convention established in our earlier work [5], in which we assume that an exponential decay model is sufficient to describe decay of CFSE in experiments for which *the first data collection time occurs after approximately 24 hours*. We note that the decay of intracellular CFSE has been observed to occur very rapidly during the first 24 hours after initial labeling and much slower thereafter [8, 35, 47] and that this observation can be fully supported with molecular-level modeling [2]. Thus, when data *are* collected in the first 24 hours, it is more accurate to describe the rate of loss of fluorescence intensity with a time-varying rate, as in a Gompertz decay model. As has been previously asserted, however, the first observation occurred approximately 24 hours after stimulation for the experimental data we consider throughout this dissertation. Therefore, we require only one parameter to completely describe the decay of CFSE: c, which is listed as parameter 3 in the Table 2.1.

We assume that each T_i^{div} has a lognormal distribution with mean and variance denoted $\mathrm{E}\left[T_i^{div}\right]$ and $\mathrm{Var}\left[T_i^{div}\right]$, respectively. We further assume that, for $i \geq 1$, all T_i^{div} are independent and identically distributed (*i.i.d.*). Such assumptions are consistent with earlier work using the cyton model [5, 28, 29], as well as experimental evidence [26]. We therefore have four parameters that completely describe $\{T_i^{div}\}$: $\mathrm{E}\left[T_0^{div}\right]$, $\mathrm{SD}\left[T_0^{div}\right] = (\mathrm{Var}\left[T_0^{div}\right])^{1/2}$, $\mathrm{E}\left[T^{div}\right] = \mathrm{E}\left[T_i^{div}\right]$ for $i \geq 1$, and $\mathrm{SD}\left[T^{div}\right] = \mathrm{SD}\left[T_i^{div}\right] = (\mathrm{Var}\left[T_i^{div}\right])^{1/2}$ for $i \geq 1$. These are listed as parameters 4 through 7, respectively.

We also assume that the random variables T_i^{die} for $i \geq 1$ are *i.i.d.* with a lognormal distribution, in this case with mean and variance denoted $\mathrm{E}\left[T^{die}\right]$ and $\mathrm{Var}\left[T^{die}\right]$, respectively.

Table 2.1: Parameters for specific mathematical model.

Number	Parameter	Description
1	$\mathrm{E}[X_a]$	mean autofluorescence
2	$\mathrm{SD}[X_a]$	std. dev. of autofluorescence
3	c	exponential decay rate for CFSE
4	$\mathrm{E}\left[T_0^{div}\right]$	mean time to divide for cells in generation $i = 0$
5	$\mathrm{SD}\left[T_0^{div}\right]$	std. dev. of time to divide for cells in generation $i = 0$
6	$\mathrm{E}\left[T^{div}\right]$	mean time to divide for cells in later generations ($i \geq 1$)
7	$\mathrm{SD}\left[T^{div}\right]$	std. dev. of time to divide for cells in later generations ($i \geq 1$)
8	$\mathrm{E}\left[T^{die}\right]$	mean time to die for cells in later generations ($i \geq 1$)
9	$\mathrm{SD}\left[T^{die}\right]$	std. dev. of time to die for cells in later generations ($i \geq 1$)
10	F_0	progressor fraction for cells in generation $i = 0$
11	D_μ	mean of discrete normal distribution (used to compute F_i for $i \geq 1$)
12	D_σ	std. dev. of discrete normal distribution (used to compute F_i for $i \geq 1$)

Again such assumptions are consistent with earlier modeling work [5, 28, 29]. We further assume that there is no death for undivided cells (i.e., those cells in generation $i = 0$). In reality, there tends to be a large die-off of cells following stimulation with PHA [10], but we reiterate that for the data analyzed in this dissertation the first measurements were made approximately 24 hours post-stimulation. As a result, *the initial conditions for our mathematical model only reflect those cells that have not died in the first 24 hours after stimulation.* Therefore, as in our earlier work [5], we assume that the cells in our "initial population" (consisting of those undivided cells that are still alive 24 hours after stimulation) that do not go on to divide experience essentially no death for the duration of the experiment. Hence, we have two parameters that completely describe $\{T_i^{die}\}$: $\mathrm{E}\left[T^{die}\right] = \mathrm{E}\left[T_i^{die}\right]$ and $\mathrm{SD}\left[T^{die}\right] = \mathrm{SD}\left[T_i^{die}\right] = (\mathrm{Var}\left[T_i^{die}\right])^{1/2}$ for $i \geq 1$, which are listed as parameters 7 and 8, respectively.

The only remaining parameters that are required to characterize our model are the progressor fractions $\{F_i\}$. We allow F_0 to be one of our required model parameters and assume that each progressor fraction F_i with $i \geq 1$ is uniquely determined by the mean and standard deviation of a "discrete normal distribution", denoted D_μ and D_σ, respectively. This is consistent with the "division destiny" approach for determining progressor fractions that has been employed by Hawkins et al. [29] and Banks et al. [5], and we refer the interested reader to the latter reference for a complete discussion of the method by which the progressor fractions are computed. We therefore have three parameters that can be used to determine all the progressor fractions: F_0, D_μ, and D_σ, which are listed as parameters 10 through 12, respectively.

Thus, our specific model depends on exactly 12 parameters. We remark that parameters 1 through 3, while important for describing the data, are not considered to be "biologically relevant" parameters in the sense that they do not have any bearing on the proliferative behavior of a population of cells. Also, parameters 4, 6, 8, and 10 are perhaps the most important of the biologically relevant parameters because their interpretation in the context of cell proliferation is the most straightforward. Note that the specific cyton-based model described here is denoted Model 6 (with exponential label decay) in our previous work [5]. Precise details of our methods for computing numerical solutions for this model are provided in Appendix B.

2.6 Statistical Model and Parameter Estimation

In order to estimate the parameters in our specific mathematical model, we must first describe a statistical model that relates observable data to the mathematical model. As was previously explained, CFSE-based flow cytometry data are typically summarized in the form of histograms, and furthermore, measured FI is commonly represented on a logarithmic scale (cf. Figure 1.4). Therefore, we begin by describing how our model can be used to obtain information on cell numbers in a form that can be compared directly with such summarized experimental data.

If we define the structured densities $\tilde{n}_i(t, \tilde{x})$ (in terms of *measured* FI) as in Section 2.2, then the structured density for the entire population of cells is

$$\tilde{n}(t, \tilde{x}) = \sum_{i=0}^{\infty} \tilde{n}_i(t, \tilde{x}) \approx \sum_{i=0}^{i_{\max}} \tilde{n}_i(t, \tilde{x}).$$

(See Section 2.1 for a discussion of how we choose $i_{\max}$ for our purposes.) Now, because CFSE histogram data are most commonly reported using a base 10 logarithmic scale, we make the change of variables $z = \log_{10}(\tilde{x})$ to obtain

$$\hat{n}(t, z) = 10^z \log(10)\, \tilde{n}(t, 10^z)$$

as the structured density in cells per base 10 log unit FI.

In the discussion that follows, we let $\vec{q}_0$ denote a hypothetical "true" parameter vector (so that, in the case of our specific mathematical model, $\vec{q}_0 \in \mathbb{R}^{12}$) and let

$$I[\hat{n}](t_j, z_k; \vec{q}_0) = \int_{z_k}^{z_{k+1}} \hat{n}(t_j, z; \vec{q}_0)\, dz \tag{2.15}$$

denote the total number of cells with log (base 10) FI in the interval $[z_k, z_{k+1}]$ at time t_j. Also, we let B denote the (fixed) total number of beads in each sample tube and b_j denote the number of beads counted for the sample measured at time t_j.

Now, let N_k^j be a random variable representing the number of cells with log FI in the interval $[z_k, z_{k+1})$ measured at time t_j. Then it has been argued [5] that

$$N_k^j \sim \mathcal{N}\left(I[\hat{n}](t_j, z_k; \vec{q}_0), \frac{B}{b_j} I[\hat{n}](t_j, z_k; \vec{q}_0) \right); \tag{2.16}$$

i.e., each N_k^j is normally distributed with mean $I[\hat{n}](t_j, z_k; \vec{q}_0)$ and variance $\frac{B}{b_j} I[\hat{n}](t_j, z_k; \vec{q}_0)$. Note that this does not lead to either (1) a constant variance model or (2) a constant coefficient of variance model. Though these latter two types of statistical models are commonly assumed to underly data-collection processes [12, 20, 43], modified residual plots indicate that (2.16) is a better choice in this case [7, 47].

Define the generalized least squares (GLS) parameter *estimator* [3, 20]

$$\vec{q}_{GLS} = \underset{\vec{q} \in \mathcal{Q}}{\operatorname{argmin}}\, J(\vec{q}; \{N_k^j\}),$$

where

$$J(\vec{q}; \{N_k^j\}) = \sum_{j,k} \frac{(I[\hat{n}](t_j, z_k; \vec{q}) - N_k^j)^2}{w_k^j}, \tag{2.17}$$

$\mathcal{Q}$ denotes the set of allowable parameter vectors, and the *weights* (selected to match the variance of the N_k^j's) are given by

$$w_k^j = \begin{cases} \frac{B}{b_j} I[\hat{n}](t_j, z_k; \vec{q}_0) & \text{for } I[\hat{n}](t_j, z_k; \vec{q}_0) > I^*, \\ \frac{B}{b_j} I^* & \text{for } I[\hat{n}](t_j, z_k; \vec{q}_0) \leq I^*. \end{cases} \tag{2.18}$$

The value of I^* is positive to prevent division by zero, and in practice it is selected such that the modified residual plots produce uniform random patterns. We once again follow the convention established in our earlier work [5] and set $I^* = 200$.

If we consider the measured data to be a set of realizations $\{n_k^j\}$ of the random variables $\{N_k^j\}$, we can obtain the GLS parameter *estimate*

$$\hat{q}_{GLS} = \underset{\vec{q} \in \mathcal{Q}}{\operatorname{argmin}} J(\vec{q}; \{n_k^j\}). \tag{2.19}$$

Note that to compute the weights $\{w_k^j\}$ we need $\vec{q}_0$, but to estimate $\vec{q}_0$ we need the weights. In order to overcome this obstacle, we use a conventional generalized least squares iterative estimation procedure [3, 12] as described in Algorithm 2.6.1. In this algorithm, note that ε is a threshold tolerance that allows the user to specify a termination criterion, $\vec{q}_{typ}$ is a vector with elements that reflect the relative sizes of the parameters to be estimated, and "./" denotes element-wise division. We provide specific details of our implementation of this algorithm in Appendix B.5.

To demonstrate the efficacy of the parameter estimation procedure, we provide sample results in Figure 2.1. This figure shows the model output for Days 1 through 5 when using the "optimal" parameter values obtained by applying Algorithm 2.6.1 to the data set depicted in Figure 1.4. We see that the mathematical model described in Section 2.5 accurately describes the overall behavior of T cells collected from healthy donors and stimulated to divide with PHA; however, as discussed by Banks et al. [5], the model does seem to include systematic errors. That is, in Figure 2.1 we see that the data is not always centered around the output of the mathematical model. Possible explanations for misspecification of the model are hinted at in Section 3.4, and are discussed further in Chapter 6. Since our statistical model does not include terms for misspecification of the mathematical model, its use to compute parameter confidence intervals might not be appropriate. We can, however, examine the reliability of the data collection process and consider variability in the parameter estimates from this perspective.

Algorithm 2.6.1 Parameter Estimation Procedure

1. Obtain initial estimate $\hat{q}^{(0)}$ using (2.19) with $w_k^j = 1$ for all j, k.
2. Compute weights w_k^j using (2.18) with $\vec{q}_0$ replaced by $\hat{q}^{(0)}$.
3. Initialize the iteration counter ℓ with the value 1.
4. Do each of the following:
   - Compute $\hat{q}^{(\ell)}$ using (2.19) with current weights w_k^j.
   - Update the weights using (2.18) with $\vec{q}_0$ replaced by $\hat{q}^{(\ell)}$.
   - Store the value of $\left|\left|[\hat{q}^{(\ell)} - \hat{q}^{(\ell-1)}]./[\vec{q}_{typ}]\right|\right|$ in Δ.
   - Increment ℓ by 1.
5. If $\Delta > \varepsilon$, return to Step 4. Otherwise, terminate the algorithm.

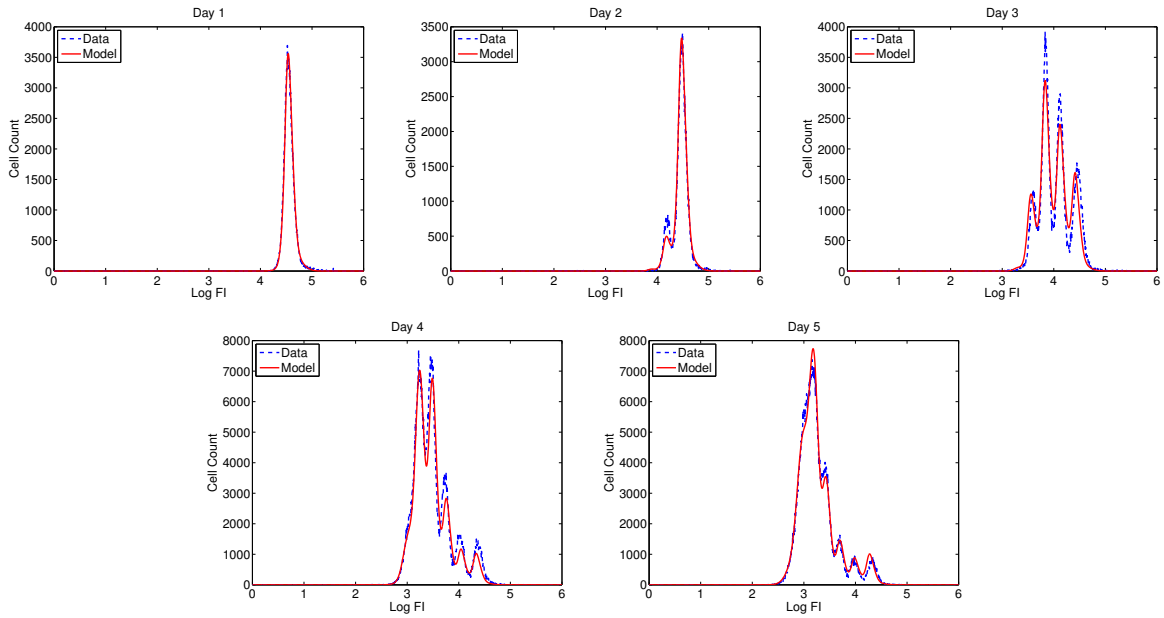


Figure 2.1: The results obtained when fitting the specific model described in Section 2.5 to data using Algorithm 2.6.1.

Chapter 3

Variability in Data and Parameter Estimates

In this chapter, we consider a large set of CFSE-based flow cytometry data that was obtained for CD4+ and CD8+ T cells collected from two healthy donors. Since triplicate measurements were recorded for each of five points in a five-day time series, we are able to analyze the experimental variability that exists in the data itself. Furthermore, the triplicate measurements make it possible to construct a considerable number of five-day data sets for each donor and cell type. In theory, each such data set should be very similar, so one might expect that a set of parameters for describing cell proliferation dynamics observed in one of the data sets should be essentially equivalent to those describing observations made using another of the data sets. By applying the parameter estimation scheme described in Section 2.6 to each five-day data set, we are able to assess experimental variability in the estimates of the parameters characterizing the specific cyton-based mathematical model described in Section 2.5. Also, by collecting data from two different human donors and considering two specific cell types, we are able to make some observations concerning biological variablility in the cell proliferation parameters. Note that many of the results described in this chapter have been previously published by the author of this dissertation and his collaborators [4, 6].

3.1 Data for Variability Study

As discussed above, the goal of the study described in this chapter is to assess the experimental and biological variability in parameter estimates produced using CFSE-based flow cytometry data and cyton-based mathematical models. To obtain such data, our experimental collaborators at Universitat Pompeu Fabra (see the "Acknowledgements" section at the beginning of this dissertation) collected blood samples from two human donors and isolated peripheral

blood mononuclear cells (PBMCs) from these samples. The PBMCs, hereafter referred to as just "cells", were then passed through a strainer to remove clumps of cells and stained with CFSE according to the standard protocol [35]. Forty-five minutes after CFSE staining, the cells were stimulated to divide by exposing them to phytohaemagglutinin (PHA), a nonspecific T cell mitogen. Then, approximately 1 million cells were placed into each of several "wells", which are typical containers for cell culture experiments. Each well contained approximately 1 mL of RPMI-1640/10% fetal calf serum (FCS), which is a typical nutrient medium for such experiments. For each one of the donors, three wells were "seeded" in this way for each of five measurement times in order to allow for measurements to be obtained in triplicate at each time point; thus, 15 wells were seeded per donor. Since two donors were considered, a total of 30 wells were seeded at the beginning of the experiment.

Here we once again note that, as cells proliferate, CFSE in a dividing mother cell is distributed to two daughter cells. The CFSE bound to the proteins inside cells also naturally degrades over time. Thus, there are two mechanisms by which the mass of CFSE per cell can decline: cell division and CFSE decay. Such a decline in CFSE per cell can be quantified by collecting cells from the wells at various points in time and passing them through a flow cytometer. So, at each of five time points corresponding to approximately 1, 2, 3, 4, and 5 days after PHA stimulation, we transferred the contents of one of the wells to a sample tube that contained a known number of "beads". (The number of beads in each sample tube is fixed and is reported by the manufacturer of the tubes.) The actual times for data collection coincided with 23.5, 46.0, 67.5, 94.5, and 117.5 hours after stimulation with PHA. In order to minimize disruption of proliferating cell populations, the contents of any given well were harvested only once; however, we remark that beginning on Day 3 (after the Day 3 measurement was made) one third of the nutrient medium in each *unused* well was exchanged with fresh medium every 24 hours and that this exchange of medium could have affected the cell cultures in the wells that were harvested after Day 3. The cells were then stained with fluorochrome-labeled antibodies (anti-CD3, anti-CD4, and anti-CD8) as well as the viability dye known as ViViD, which allows for the identification of dead cells that haven't yet disintegrated. A sample consisting of a fraction (about 10 to 50%) of the contents of the sample tube was then passed through a flow cytometer. During this process, the flow cytometer measured the fluorescence intensity (FI) at various wavelengths of each cell in the sample and counted the number of beads in the sample. Because we processed only a fraction of the contents of each tube, the "actual" cell counts for any given range of FI were obtained by scaling the observed counts upward by the ratio of known number of beads in the tube to counted beads for the sample. As mentioned above, this process was repeated using cells from three different wells (per donor) at each measurement time.

Because the FI emitted at wavelengths in the range 515 to 545 nm varies directly with the

mass of CFSE within a cell [35, 48], this FI is a useful surrogate for CFSE mass. Also, because CFSE is allocated evenly (by assumption) to two daughter cells upon cell division, FI histograms associated with a population of cells acquire more "peaks" as the cells divide asynchronously. In Figures 3.1 through 3.8, we present summary histograms for the data collected in our study. Each of these figures illustrates the preceding point.

As described in the preceding paragraph, the FI emitted in the "CFSE range" by a given cell may be taken to be synonymous with the mass of CFSE contained in that cell. Therefore, CFSE data can be used to validate a mathematical model describing cell population dynamics that is based upon mass conservation principles. Such a model is described in detail in Chapter 2. Furthermore, when cells have been labeled with other markers (as is the case for the experiments in this study), information about FI at other wavelengths can be used to distinguish different types of cells (e.g., CD4+ versus CD8+ T cells, or living versus dead cells) [35].

In order to ascertain variability in parameter estimates, we require a considerable number of data sets. To this end, we use various combinations of the triplicate samples collected on Days 1 through 5 to form a large number of time series data sets. Since three samples were collected on each of five days, there should be $3^5 = 243$ possible ways to form a five-day "longitudinal" data set for each donor. One such five-day data set is depicted in Figure 3.9. We remark that data for one of the samples corresponding to Donor 1 and Day 4 was not available due to a data collection error; therefore, there are in fact only $3^4 \times 2 = 162$ possible ways to form a five-day data set for Donor 1. It should be explicitly noted that data sets formed in this way *do not represent truly longitudinal data* because measurements corresponding to each time point were made using distinct cell cultures (wells). In this type of *in vitro* experiment [5, 29, 35, 39], *it is tacitly assumed that the populations of cells in each well are identical* (up until the moment cells are harvested from a particular well) in that they include the same numbers of total cells in the same proportions (according to cell type). This assumption allows one to interpret time series data sets formed as described above as having come from longitudinal observations. In practice, however, there can be considerable variation in the cell cultures in the various wells due to experimental error in the initial seeding of the wells (among other reasons). This issue will be discussed further in Section 3.2.

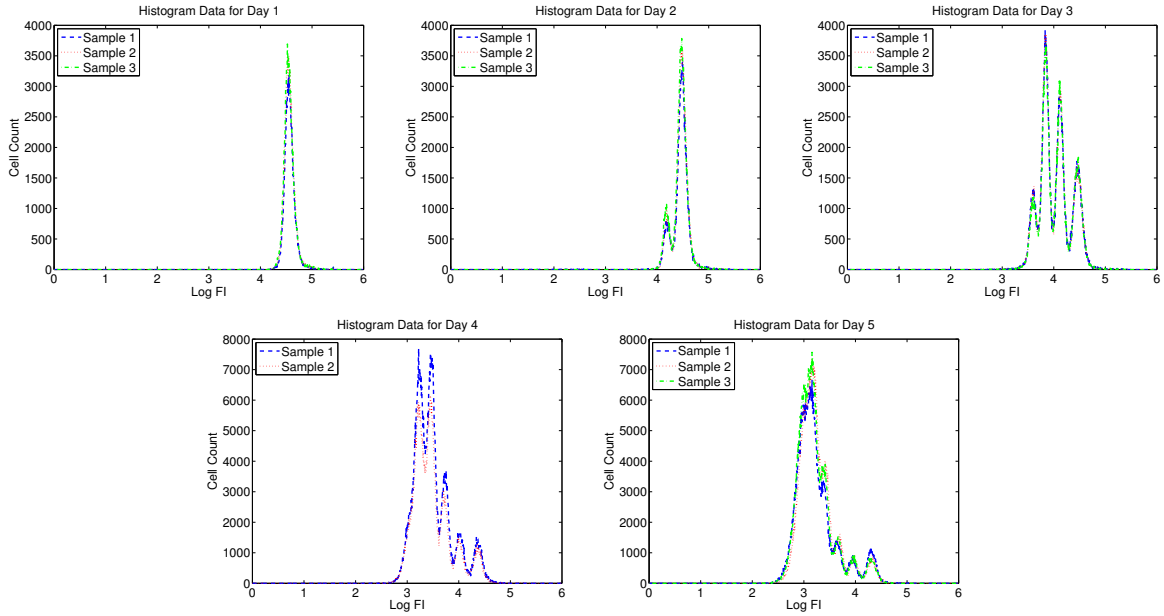


Figure 3.1: Summary histogram data for CD4+ T cells measured for Donor 1 using ViViD dye to exclude dead cells.

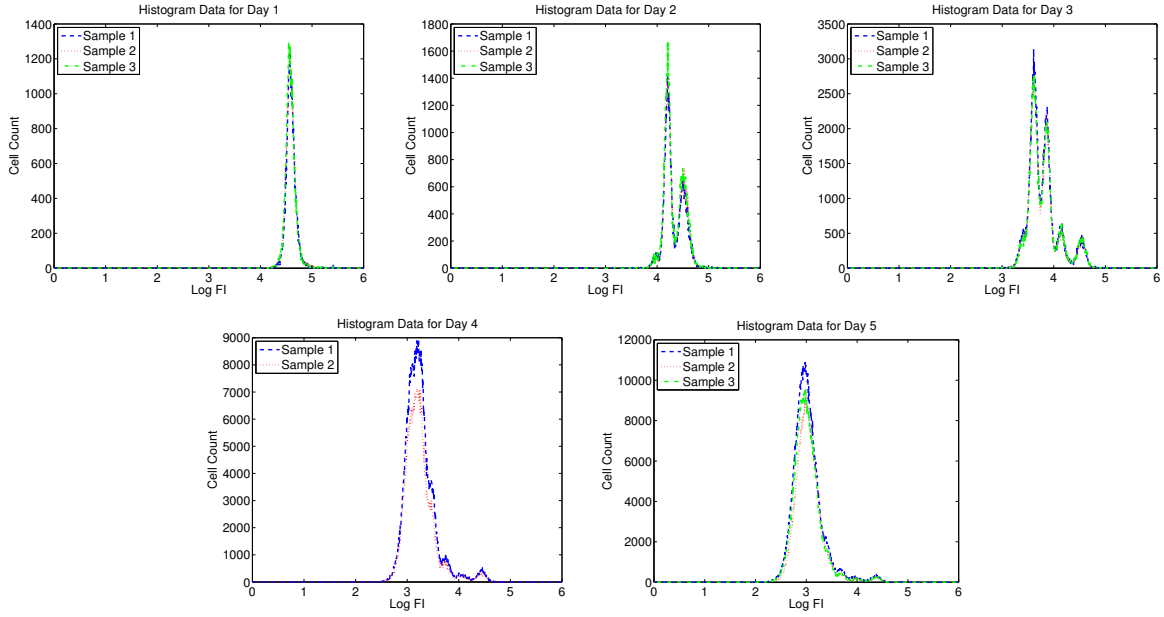


Figure 3.2: Summary histogram data for CD8+ T cells measured for Donor 1 using ViViD dye to exclude dead cells.

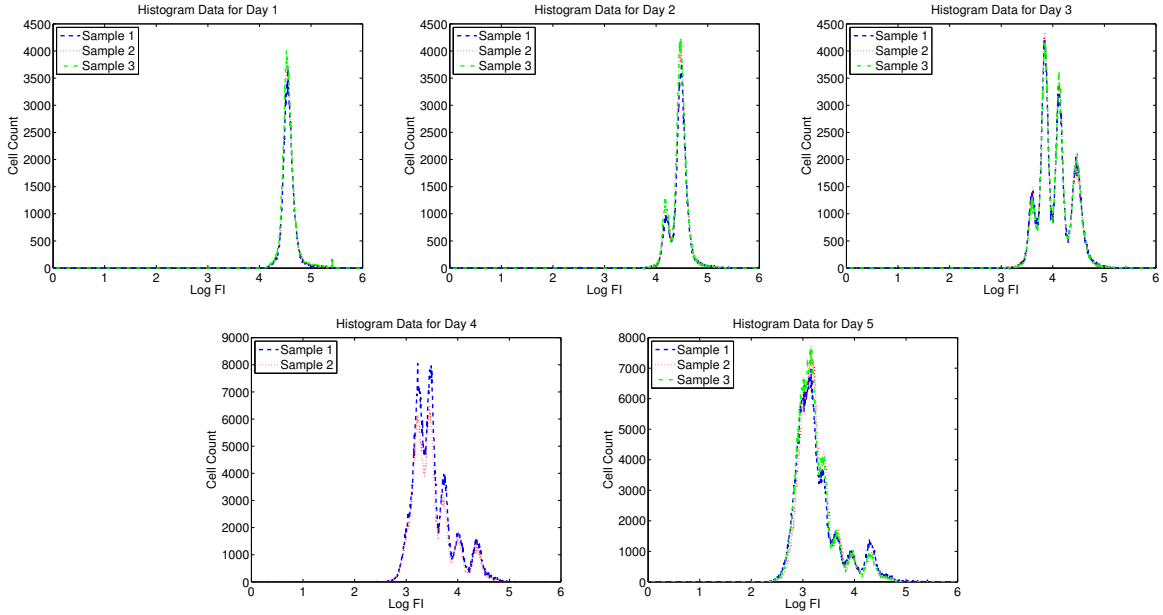


Figure 3.3: Summary histogram data for CD4+ T cells measured for Donor 1 without using ViViD dye to exclude dead cells.

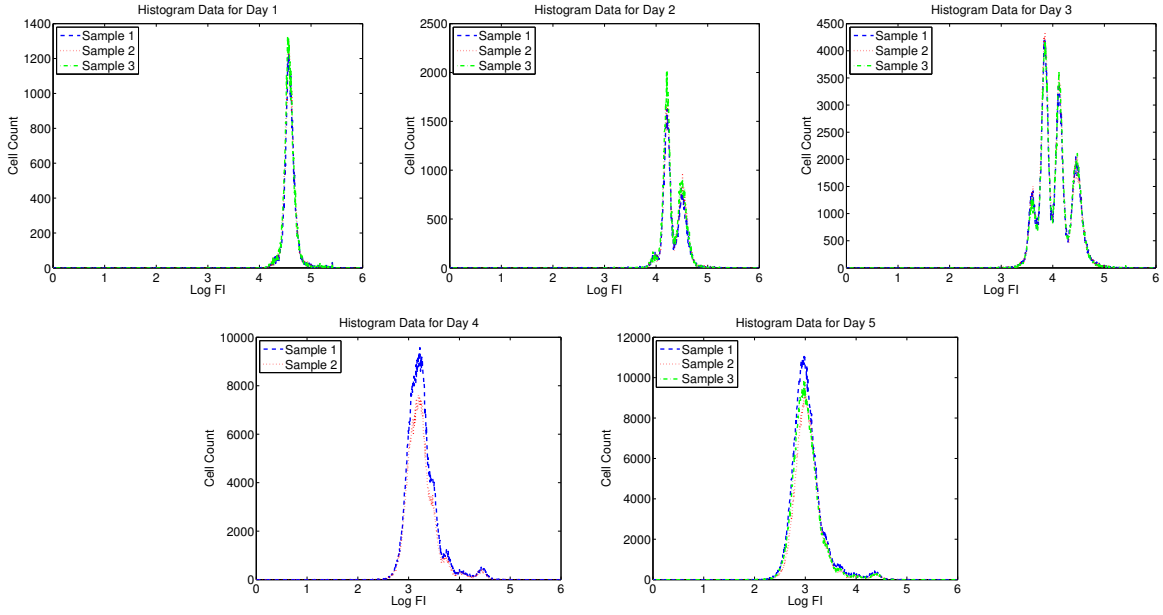


Figure 3.4: Summary histogram data for CD8+ T cells measured for Donor 1 without using ViViD dye to exclude dead cells.

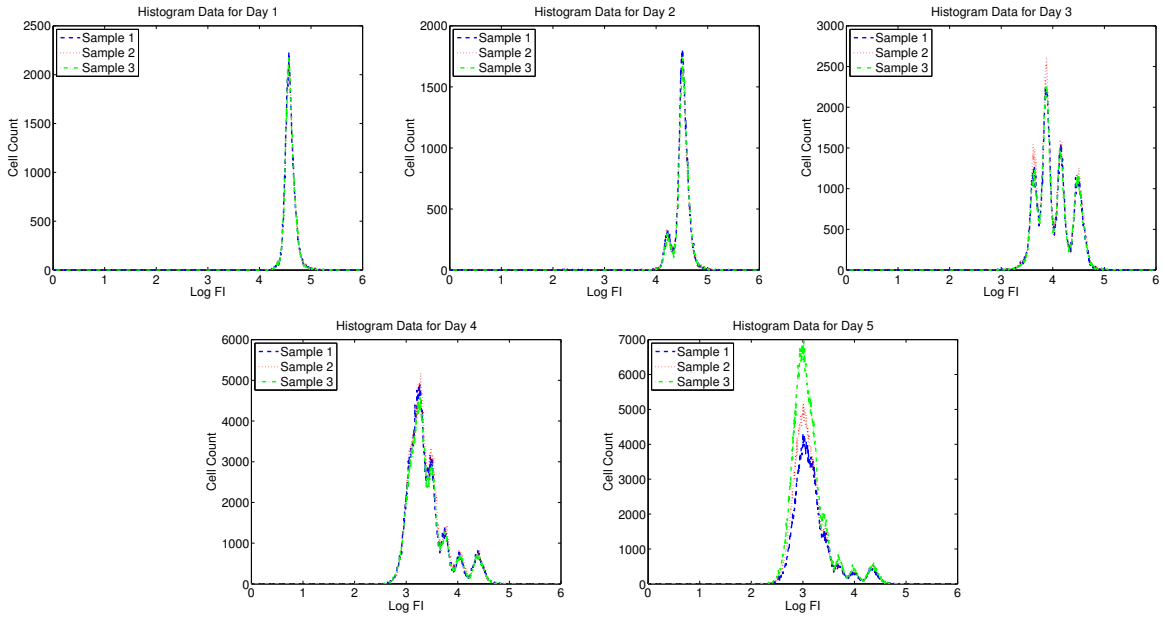


Figure 3.5: Summary histogram data for CD4+ T cells measured for Donor 2 using ViViD dye to exclude dead cells.

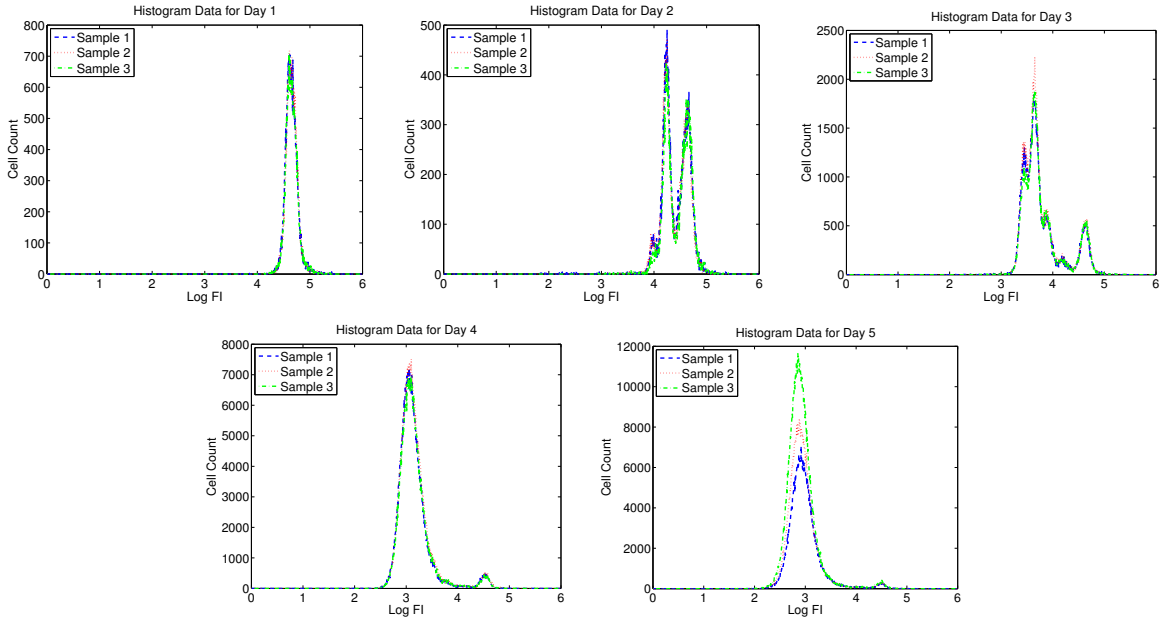


Figure 3.6: Summary histogram data for CD8+ T cells measured for Donor 2 using ViViD dye to exclude dead cells.

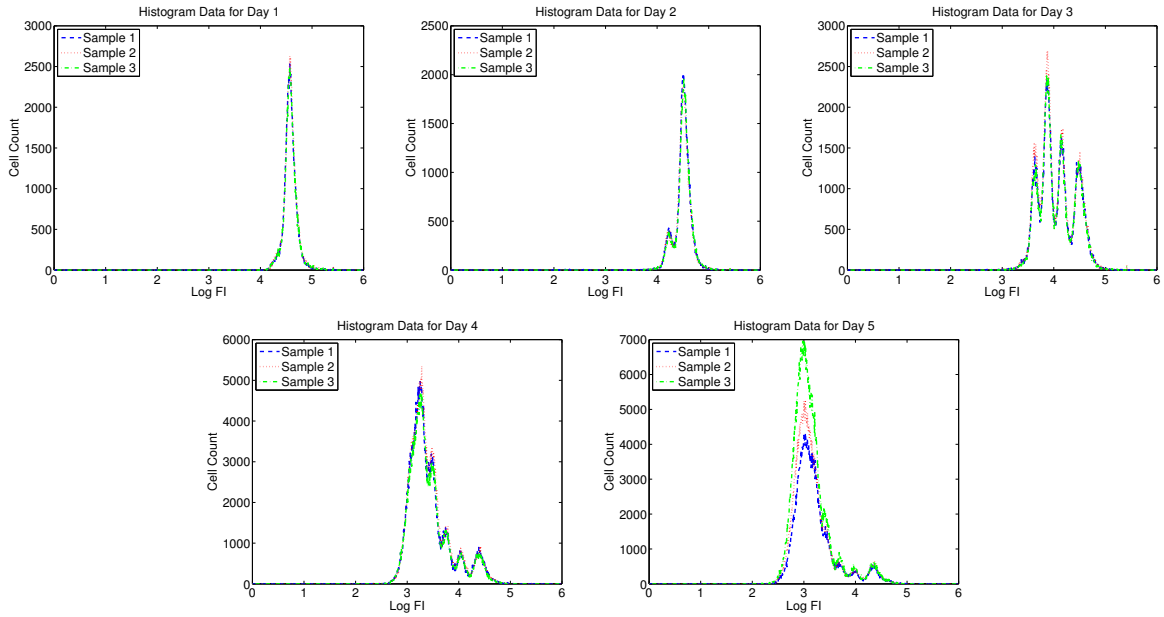


Figure 3.7: Summary histogram data for CD4+ T cells measured for Donor 2 without using ViViD dye to exclude dead cells.

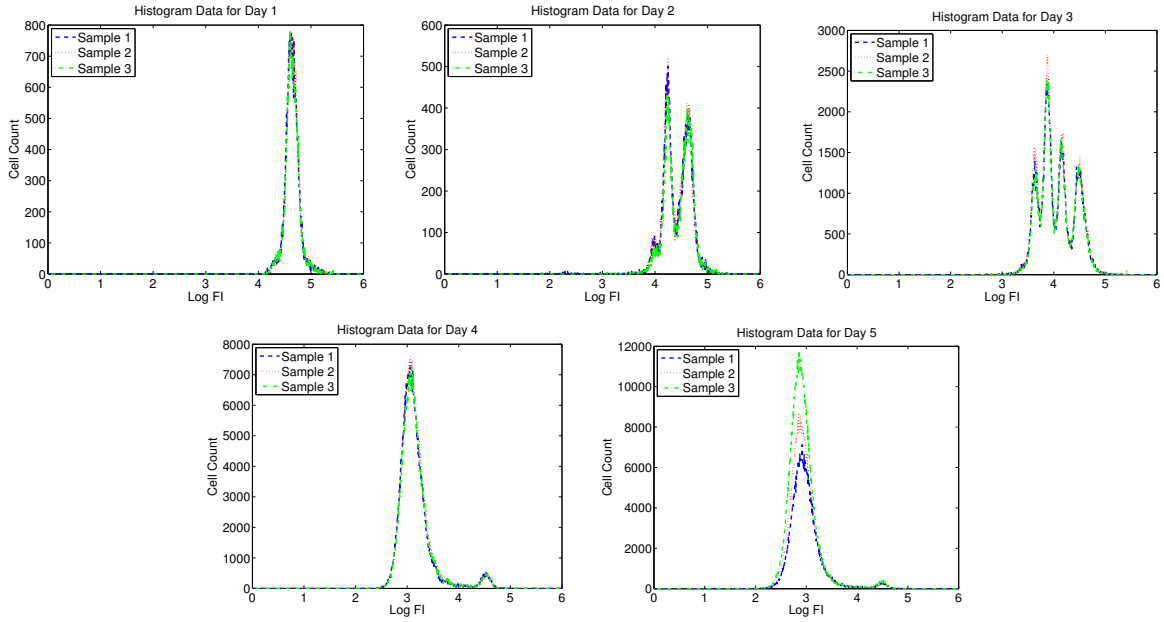


Figure 3.8: Summary histogram data for CD8+ T cells measured for Donor 2 without using ViViD dye to exclude dead cells.

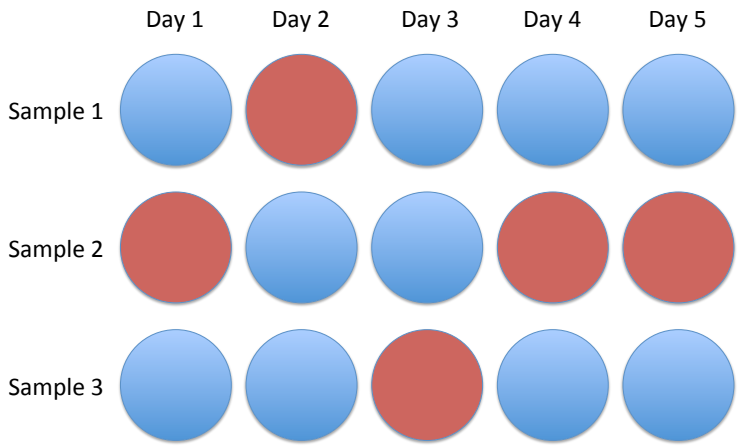


Figure 3.9: Schematic showing the wells to be used for triplicate measurements on each of five days. The five wells distinguished by the (darker) color red can be used to form one of 243 possible five-day data sets.

3.2 Variability in the Data

It is clear from Figures 3.1 through 3.8 that, while the *shapes* of the histograms summarizing observed data do not vary much between triplicate experiments for any particular day and combination of donor, ViViD dye status, and cell type, the *scale* of the histograms can vary considerably. That is, there can be a significant difference between triplicate experiments in the *number* of cells represented in the respective histograms. One can also see in Figures 3.1 through 3.8 that there seems to be more variability in the cell counts at the later time points (those corresponding to Day 4 and especially those corresponding to Day 5). Tables 3.1 through 3.5 list the cell counts for each sample and each combination of donor, ViViD dye status, and cell type on Days 1 through 5, respectively, and can be compared with the histograms in Figures 3.1 through 3.8. For example, Table 3.1 shows all the cell counts in the various samples represented in the upper-left (Day 1) plots for each of Figures 3.1 through 3.8. Tables 3.1 through 3.5 also show an estimate of the coefficient of variation based on each row of sample cell counts. The coefficient of variation is a measure of the relative variation in the cell counts, and can be estimated as

$$\hat{c}_v = \frac{s}{\bar{x}},$$

where $\bar{x}$ denotes the sample mean of the cell counts (in the relevant biological samples) and s denotes the sample standard deviation. We refer to $\hat{c}_v$ as an "estimate" because it is derived from a finite "sample" of data (in our case, from three cell counts derived from three biological

samples). The true coefficient of variation would be computed as

$$c_v = \frac{\sigma}{\mu},$$

where μ and σ are the mean and standard deviation, respectively, of cell counts for the "population" of all possible biological samples for the specific time, donor, ViViD status, and cell type in question.

In order to understand how variation in the cell counts arises, recall from Section 3.1 that *approximately* 1 million cells were placed into each of several wells, and that each distinct sample was drawn from a distinct well. Because of the experimental error inherent in seeding the wells, we expect that each well actually started out with a different number of cells. We should remark here that the numbers in Tables 3.1 through 3.5 do not represent total cell counts, but rather they represent cell counts for specific types of cells (CD4+ or CD8+ T cells) under specific conditions (donor and ViViD dye status). So, for example, if we attempt to seed 1 million cells from Donor 1 into a well and the true proportion of CD4+ T cells is 15% for this donor, there should be about 150 thousand CD4+ T cells in the well at time $t = 0$; however, there will be some variation in this number (150 thousand) because of the variation in the total cell population number (1 million). We did not make any measurements at time $t = 0$, so we cannot directly assess the variation present in the numbers of cells initially seeded into the wells. Our best approximation of this initial variation comes from the cell counts observed on Day 1, which are shown in Table 3.1. Also, because the true proportion of cells (out of approximately 1 million) corresponding to a particular day, donor, ViViD dye status, and cell type varies, we cannot directly compare all 24 cell count numbers in one of the tables, and we cannot directly compare the 8 *sample variances* or *sample standard deviations* obtained for the 8 rows in any given table. To be clear, we cannot use such direct comparisons because the *magnitudes* of the cell count numbers tend to be different in each row of a given table. We can, however, compare some measure of *relative* variation for each of the rows of a given table. The coefficient of variation described previously is one such measure. So, the 8 numbers listed in the last column of Table 3.1 give some indication of the variability we expect when attempting to seed 1 million cells into a well, and, importantly, they can be compared with one another.

If we assume that the seeding of 1 million cells into a well is a process that is subject to random error, then the amount of error is a random variable with some well-defined probability distribution. For our purposes, the "amount of error in the initial seeding" is synonymous with the "amount of relative variation in the cell counts at Day 1", which we choose to measure using the coefficients of variation described previously. One important question to consider, then, is whether or not the amount of relative variation in cell counts (or more precisely, the sampling distribution of this statistic) changes between measurement times (days). As was

Table 3.1: Status of cell cultures at Day 1. The numbers in the columns corresponding to Samples 1, 2, and 3 represent cell counts after scaling (using bead counts). Each number in the C.V. column represents the coefficient of variation for the cell counts in the corresponding row.

Donor	ViViD Used	Cell Type	Sample 1	Sample 2	Sample 3	C.V.
1	Y	CD4	104124	117820	122408	8.29%
1	Y	CD8	36778	38924	40190	4.47%
1	N	CD4	128161	140770	146369	6.74%
1	N	CD8	39899	42259	43478	4.34%
2	Y	CD4	68651	68439	68680	0.19%
2	Y	CD8	28699	28574	27387	2.57%
2	N	CD4	91542	95914	90740	3.00%
2	N	CD8	33983	34605	32692	2.89%

Table 3.2: Status of cell cultures at Day 2. The numbers in the columns corresponding to Samples 1, 2, and 3 represent cell counts after scaling (using bead counts). Each number in the C.V. column represents the coefficient of variation for the cell counts in the corresponding row.

Donor	ViViD Used	Cell Type	Sample 1	Sample 2	Sample 3	C.V.
1	Y	CD4	128535	139264	149408	7.51%
1	Y	CD8	58115	61406	67116	7.32%
1	N	CD4	156211	171078	181924	7.61%
1	N	CD8	70147	75010	82658	8.31%
2	Y	CD4	66884	63458	63162	3.21%
2	Y	CD8	29242	28221	26734	4.49%
2	N	CD4	82968	80701	80098	1.86%
2	N	CD8	33223	33107	31048	3.77%

Table 3.3: Status of cell cultures at Day 3. The numbers in the columns corresponding to Samples 1, 2, and 3 represent cell counts after scaling (using bead counts). Each number in the C.V. column represents the coefficient of variation for the cell counts in the corresponding row.

Donor	ViViD Used	Cell Type	Sample 1	Sample 2	Sample 3	C.V.
1	Y	CD4	278399	277201	269793	1.69%
1	Y	CD8	190473	183328	179769	2.95%
1	N	CD4	331038	338423	328589	1.54%
1	N	CD8	232513	228392	228203	1.06%
2	Y	CD4	187718	206164	184355	6.09%
2	Y	CD8	137646	150831	136055	5.73%
2	N	CD4	212391	229784	208054	5.31%
2	N	CD8	145740	158148	143890	5.19%

Table 3.4: Status of cell cultures at Day 4. The numbers in the columns corresponding to Samples 1, 2, and 3 represent cell counts after scaling (using bead counts). Each number in the C.V. column represents the coefficient of variation for the cell counts in the corresponding row.

Donor	ViViD Used	Cell Type	Sample 1	Sample 2	Sample 3	C.V.
1	Y	CD4	746988	611600	N/A	14.09%
1	Y	CD8	778452	616165	N/A	16.46%
1	N	CD4	813670	672320	N/A	13.45%
1	N	CD8	845757	672287	N/A	16.16%
2	Y	CD4	465445	495995	450620	4.92%
2	Y	CD8	562295	596485	550758	4.17%
2	N	CD4	483623	512917	466908	4.77%
2	N	CD8	572560	606627	560635	4.12%

Table 3.5: Status of cell cultures at Day 5. The numbers in the columns corresponding to Samples 1, 2, and 3 represent cell counts after scaling (using bead counts). Each number in the C.V. column represents the coefficient of variation for the cell counts in the corresponding row.

Donor	ViViD Used	Cell Type	Sample 1	Sample 2	Sample 3	C.V.
1	Y	CD4	675662	703755	731910	4.00%
1	Y	CD8	960299	763048	823195	11.91%
1	N	CD4	755566	752181	780992	2.06%
1	N	CD8	998167	794116	853743	11.90%
2	Y	CD4	418710	500372	659604	23.28%
2	Y	CD8	578203	700480	940572	24.92%
2	N	CD4	435902	519138	678801	22.67%
2	N	CD8	590954	712185	954256	24.58%

previously asserted, Figures 3.1 through 3.8 provide convincing evidence that the amount of relative variation in cell counts does change with respect to time. More specifically, these figures lead us to suspect that there is a significant difference between the relative variation observed in the earlier days of the experiment (1 through 3) and that observed in the later days (4 and 5). The coefficient of variation estimates in the last columns of Tables 3.1 through 3.5 can be used to demonstrate this claim conclusively through the use of formal statistical hypothesis testing.

The two-sided Wilcoxon rank-sum test allows one to determine if two independent samples have been drawn from the same continuous distribution [31]. In our case, we assume that the eight coefficient of variation numbers corresponding to Day i are eight realizations (constituting a sample) of a continuous random variable with cumulative density function (cdf) F_i for $i \in \{1, 2, 3, 4, 5\}$ and that each of the five samples is independent of all of the others. Thus, for any particular pair of days (i, j), we would like to test the null hypothesis that the corresponding

samples of coefficient of variation numbers were drawn from the same population of values (or, equivalently, from two populations with identical cdfs); i.e.,

$$H_0 : F_i(x) = F_j(x) \text{ for all } x.$$

For the Wilcoxon test, the alternative hypothesis is that one of the two samples was drawn from a population that tends to have larger values than the population from which the other sample was drawn (or, equivalently, that one of the two corresponding distributions is *stochastically larger* than the other); i.e.,

$$\begin{aligned} H_A : \quad & F_i(x) \leq F_j(x) \text{ for all } x, \text{ with strict inequality for at least some } x, \\ & \text{or } F_i(x) \geq F_j(x) \text{ for all } x, \text{ with strict inequality for at least some } x. \end{aligned}$$

As with any statistical hypothesis test, the Wilcoxon test produces a "test statistic" that can be converted into a "p-value". The p-value indicates the probability of obtaining a test statistic at least as extreme as the one which was actually observed *assuming that the null hypothesis is true*. If we use a "significance level" of 0.05, then we are asserting that outcomes with probability less than 0.05 are unlikely to occur. Therefore, we should reject the null hypothesis whenever the p-value is less than 0.05.

Suppose, for example, that we want to test the claim that there is no difference between the distribution of relative variations in cell counts observed at Day 1 and that observed at Day 2. This null hypothesis can be formalized as

$$H_0 : F_1(x) = F_2(x) \text{ for all } x,$$

where F_1 and F_2 denote the cdfs for the coefficients of variation in cell counts at Days 1 and 2, respectively. The alternative hypothesis is that one of the two distributions is *stochastically larger* than the other; i.e.,

$$\begin{aligned} H_A : \quad & F_1(x) \leq F_2(x) \text{ for all } x, \text{ with strict inequality for at least some } x, \\ & \text{or } F_1(x) \geq F_2(x) \text{ for all } x, \text{ with strict inequality for at least some } x. \end{aligned}$$

We can use the coefficient of variation estimates from Tables 3.1 and 3.2 to perform a Wilcoxon test for these hypotheses. That is, we can compare the eight coefficient of variation numbers from Table 3.1 with the eight coefficient of variation numbers from Table 3.2 using a Wilcoxon test. Based on the p-value of 0.2345 that results from performing this test, we fail to reject the null hypothesis (using a 0.05 significance level) and conclude that there is not a statistically significant difference between the relative variations in cell counts observed at Days 1 and 2,

respectively. Similarly, we can perform Wilcoxon tests to conclude that there is not a statistically significant difference between the relative variations in cell counts observed at Days 1 and 3 ($p = 0.8785$) or the relative variations in cell counts observed at Days 2 and 3 ($p = 0.1304$). On the other hand, Wilcoxon tests lead us to conclude that there *is* a statistically significant difference between the relative variations in cell counts observed at Days 1 and 4 ($p = 0.03792$) and between the relative variations in cell counts observed at Days 1 and 5 ($p = 0.02813$). Similarly, there appears to be a statistically significant difference between the relative variations in cell counts observed at Days 2 and 5 ($p = 0.03792$) and between the relative variations in cell counts observed at Days 3 and 5 ($p = 0.01476$). (Wilcoxon tests do not indicate a significant difference between relative variations in cell counts observed at Days 2 and 4 or Days 3 and 4, but recall that at least some of the Day 4 coefficients of variation are based on fewer cell counts because fewer biological samples were used for Donor 1 on Day 4; therefore, information concerning variation in cell counts for Day 4 is considerably less reliable.)

We propose that this change in the relative variation of the cell counts with respect to time cannot be explained by proliferation dynamics alone. In fact, if we assume that our mathematical model describing cell proliferation is correct, the relative variation in cell counts for distinct cultures proliferating with the same dynamics should not change in time. Relative variation can be measured in terms of percent differences or coefficients of variation and we offer proofs of the assertion in the preceding sentence with respect to both of these measures of relative variation in Appendix A.3, but it is easy (and instructive) to understand the validity of the assertion under the assumption of a simple exponential growth model. So, consider two cultures of cells, "A" and "B", that are proliferating at the same exponential growth rate α. If A_0 and B_0 denote the initial numbers of cells present in cultures A and B, respectively, then

$$A(t) = A_0 e^{\alpha t}$$

and

$$B(t) = B_0 e^{\alpha t}$$

describe the numbers of cells in the respective cultures at time t. The initial percent difference in the cell counts is

$$\frac{2(A_0 - B_0)}{A_0 + B_0},$$

while the percent difference at any later time t is

$$\frac{2(A_0 e^{\alpha t} - B_0 e^{\alpha t})}{A_0 e^{\alpha t} + B_0 e^{\alpha t}} = \frac{2(A_0 - B_0)e^{\alpha t}}{(A_0 + B_0)e^{\alpha t}} = \frac{2(A_0 - B_0)}{A_0 + B_0}.$$

Therefore, the percent difference in cell counts does not change with respect to time. Note that

the amount of statistical *variation* (e.g., difference) in the cell counts *does* generally change in time, but the *relative variation* (e.g., percent difference) must remain constant.

Since we have clearly demonstrated that the relative variation in the cell counts does, indeed, change between Days 1, 2, and 3 and Day 5, then some source of variability must exist which is not accounted for in our model. We propose that the exchange of nutrient medium starting at Day 3 could be such a source of variability. Exchange of nutrient medium could feasibly remove, disturb, or damage some of the cells in the affected wells, and it would certainly change the amount of nutrient available to the cells in those wells. In fact, *the very reason that the nutrient medium is replenished starting at Day 3 is that by that time it has begun to change color, indicating that the nutrient levels have declined.* So, in addition to the changes in the amount of nutrient available to the cell cultures that occur at discrete points in time corresponding to nutrient medium exchange, we may infer that cells growing and dividing in the various cell cultures significantly deplete nutrients in their respective wells *throughout the experiment.*

3.3 Variability in the Parameter Estimates

In order to assess variability in parameter estimates, we applied the parameter estimation technique described in Section 2.6 to the various five-day time series data sets described in Section 3.1. For each of the 12 parameters from our specific mathematical model and each of the eight combinations of donor ("Donor1" or "Donor2"), ViViD dye status ("Vivid" indicating that ViViD dye was used to exclude dead cells or "NoVivid" indicating otherwise), and cell type ("CD4" or "CD8"), this led to either 162 or 243 (depending on donor, cf. Section 3.1) parameter estimates. Each such set of parameter estimates can be represented by a box plot, so for each model parameter we can construct eight box plots as illustrated in Figures 3.10 through 3.21. The box plots in each of these figures adhere to the following conventions: (i) the median value is indicated by a red line; (ii) the first and third quartiles (Q_1 and Q_3) are indicated by the lower and upper boundaries of the blue box, respectively; (iii) any value that falls above $Q_3 + 1.5(Q_3 - Q_1)$ or below $Q_1 - 1.5(Q_3 - Q_1)$ is considered to be an outlier, and is indicated by a "+"; and (iv) the black horizontal lines above and below the box (which are in most cases connected to the box via dashed vertical lines) represent the maximum and minimum values, respectively, excluding outliers.

Sets of box plots such as those described above can provide a wealth of information concerning the variability in parameter estimates and identifiability of the corresponding parameters. Individually, each box plot can be used to determine a median parameter estimate and to visualize the variation (spread) in parameter estimates for a given donor and cell type when multiple five-day data sets are considered. The amount of spread in each box plot can also be used to conclude whether or not a particular parameter is likely to be identifiable for any particular

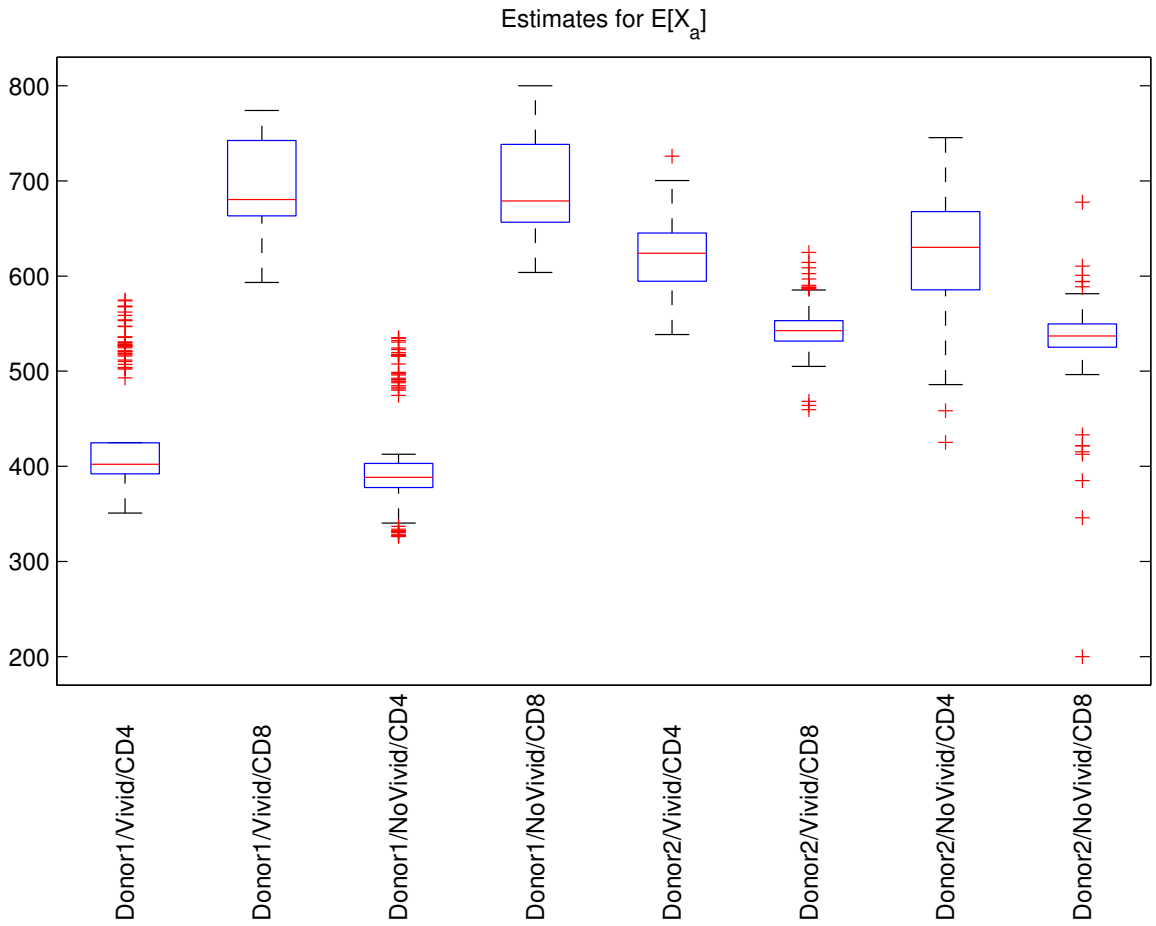


Figure 3.10: Box plots illustrating variability in estimates for the parameter $\mathrm{E}[X_a]$.

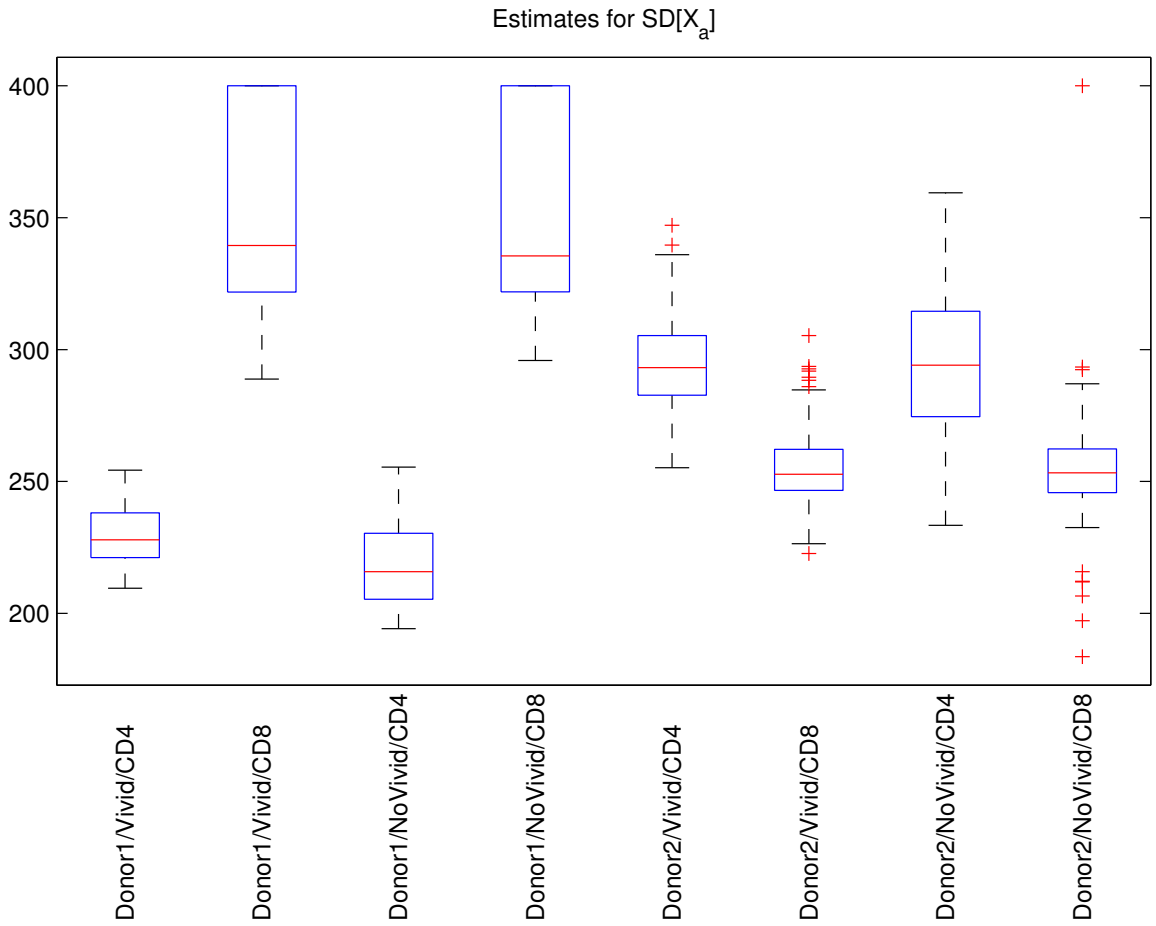


Figure 3.11: Box plots illustrating variability in estimates for the parameter SD $[X_a]$.

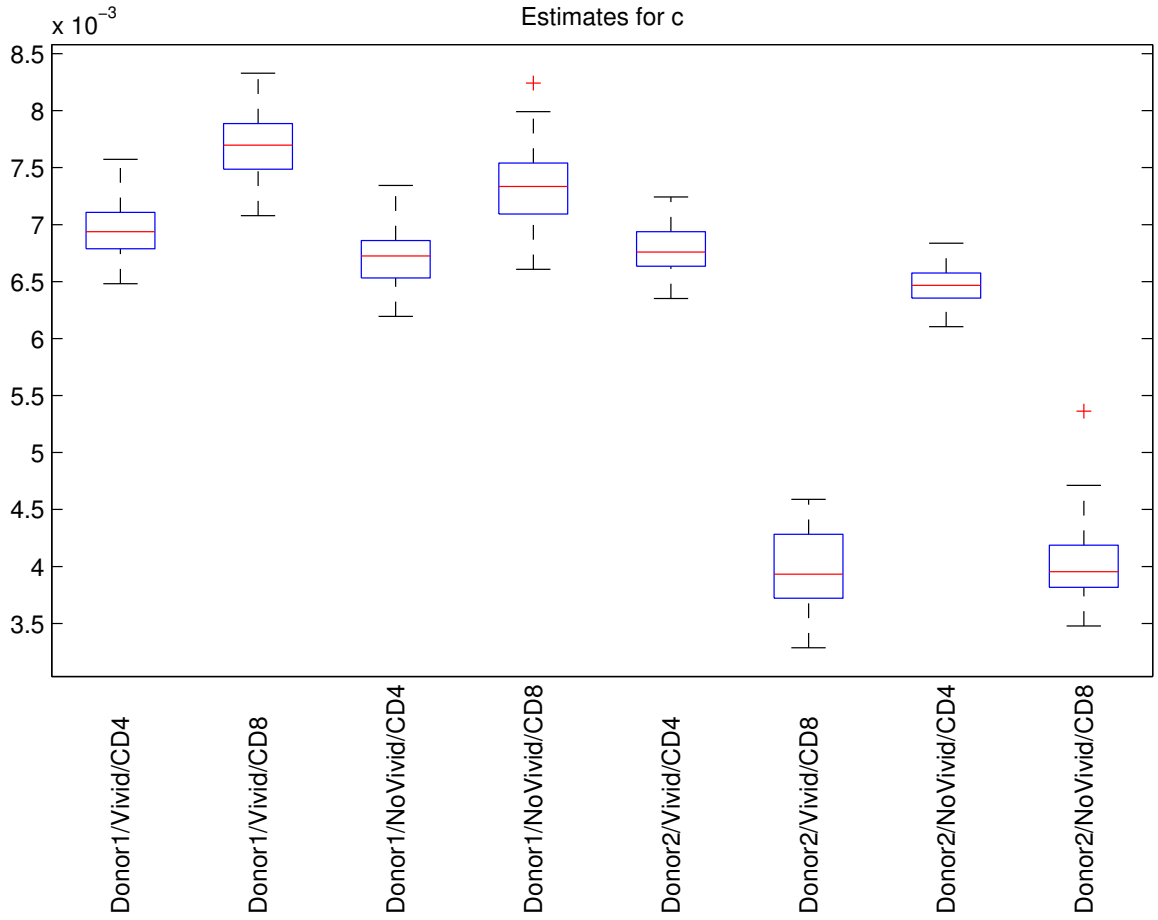


Figure 3.12: Box plots illustrating variability in estimates for the parameter c.

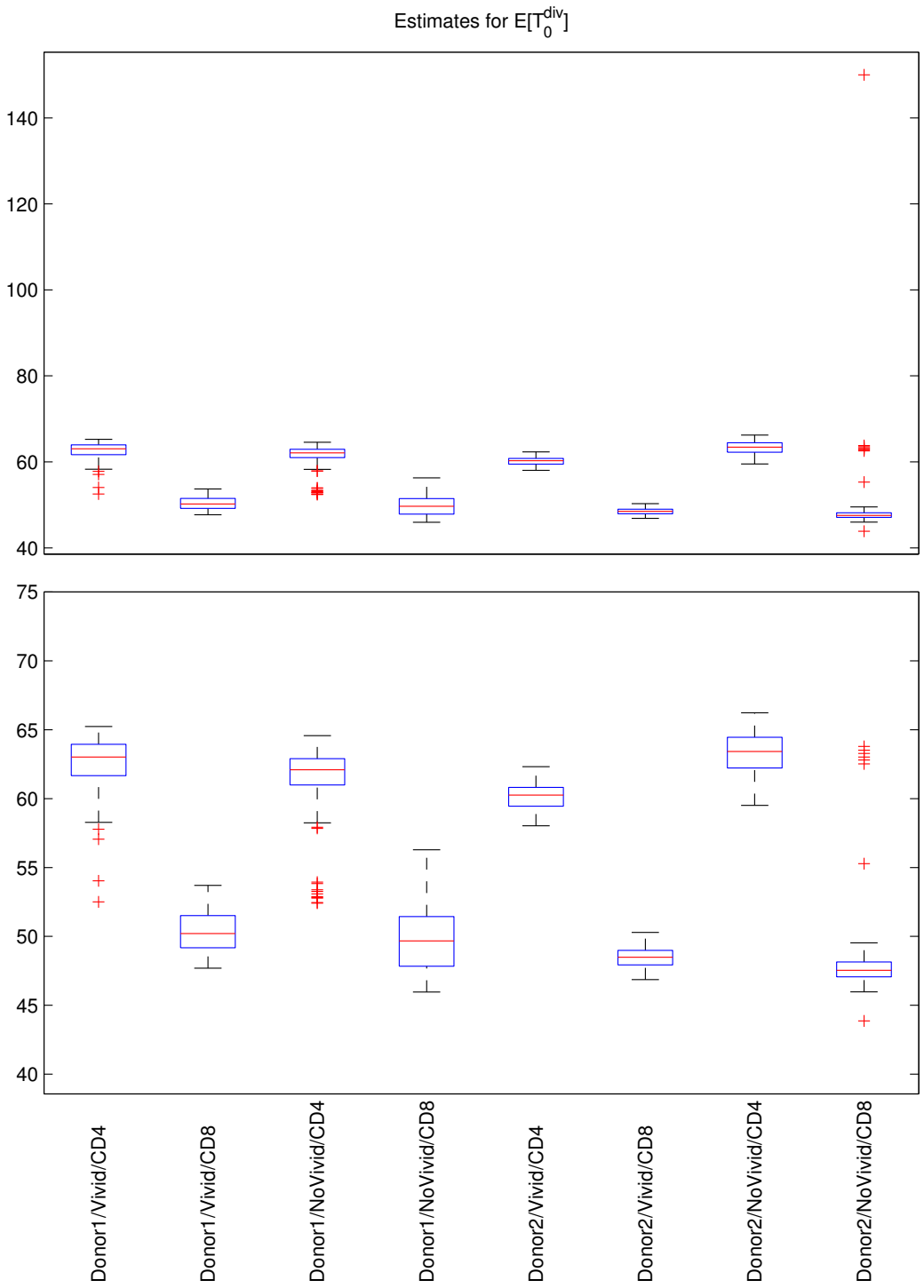


Figure 3.13: Box plots illustrating variability in estimates for the parameter E $\left[T_0^{div}\right]$. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

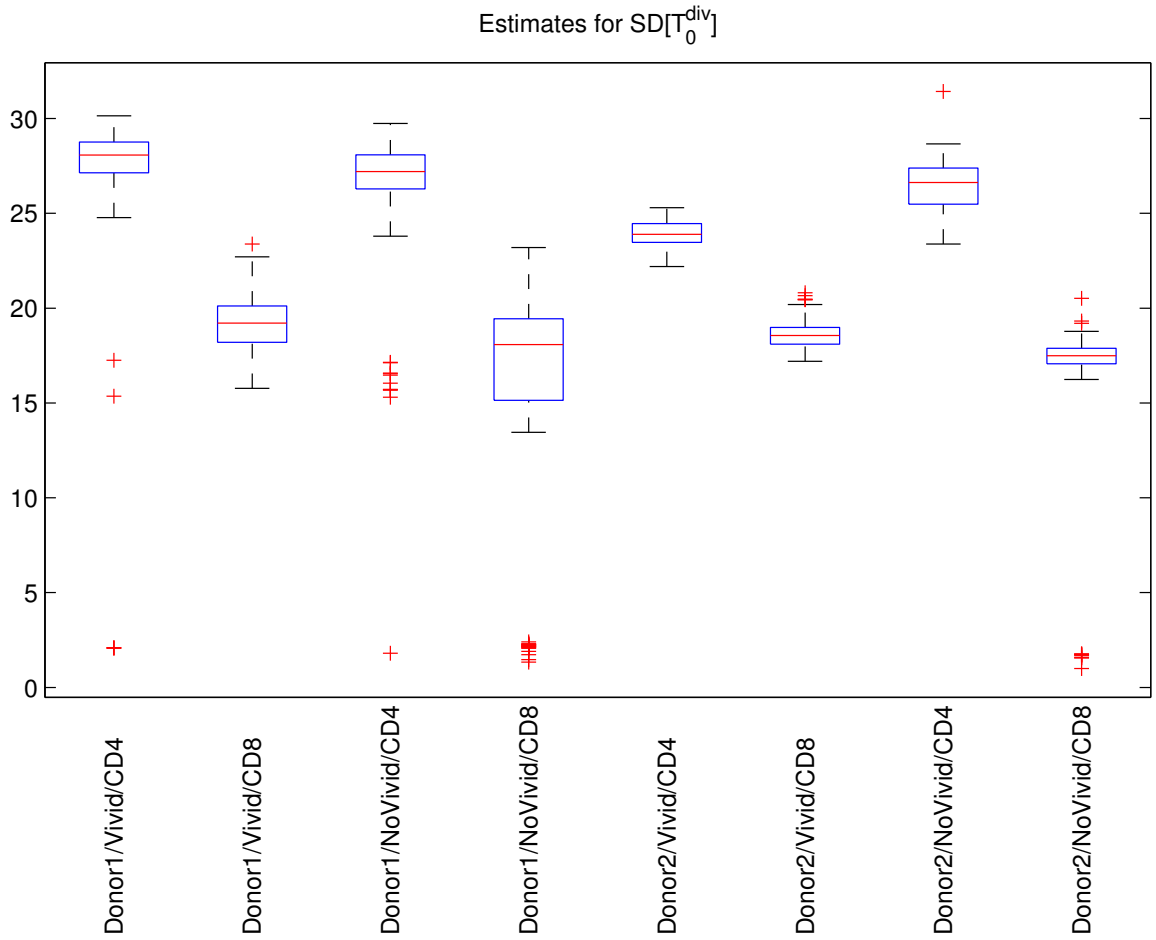


Figure 3.14: Box plots illustrating variability in estimates for the parameter SD $\left[T_0^{div}\right]$.

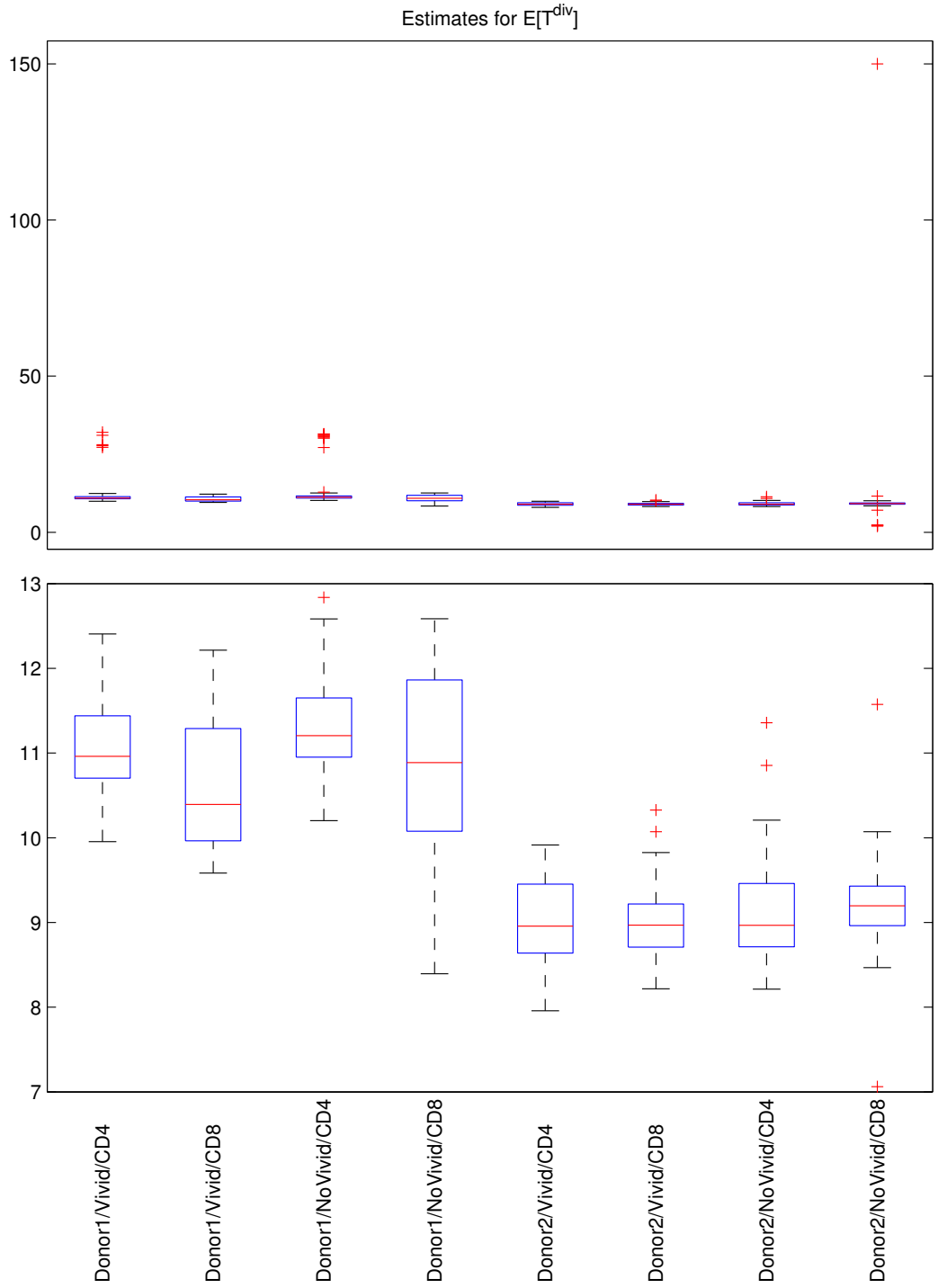


Figure 3.15: Box plots illustrating variability in estimates for the parameter E $\left[T^{div}\right]$. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

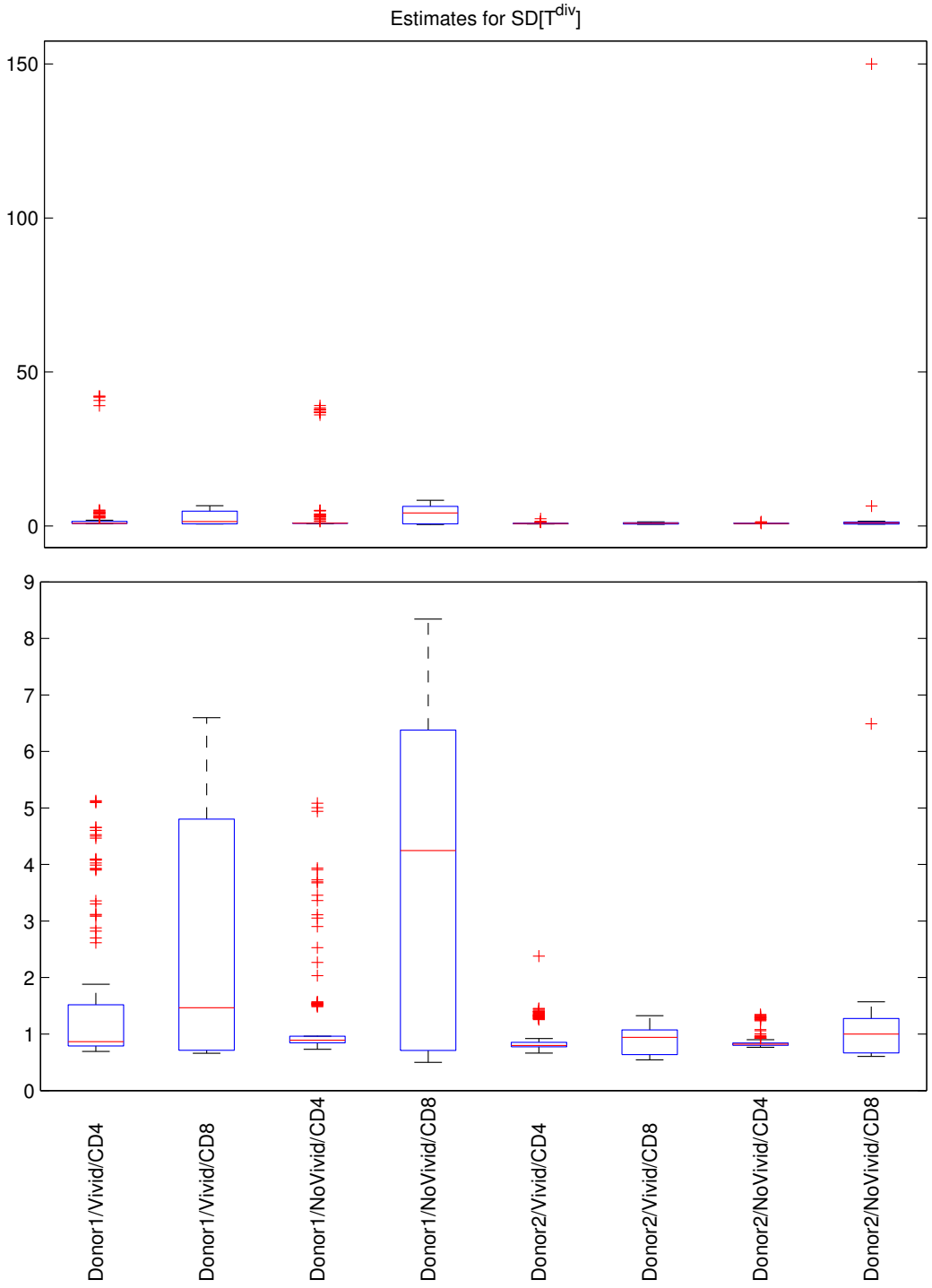


Figure 3.16: Box plots illustrating variability in estimates for the parameter SD $\left[T^{div}\right]$. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

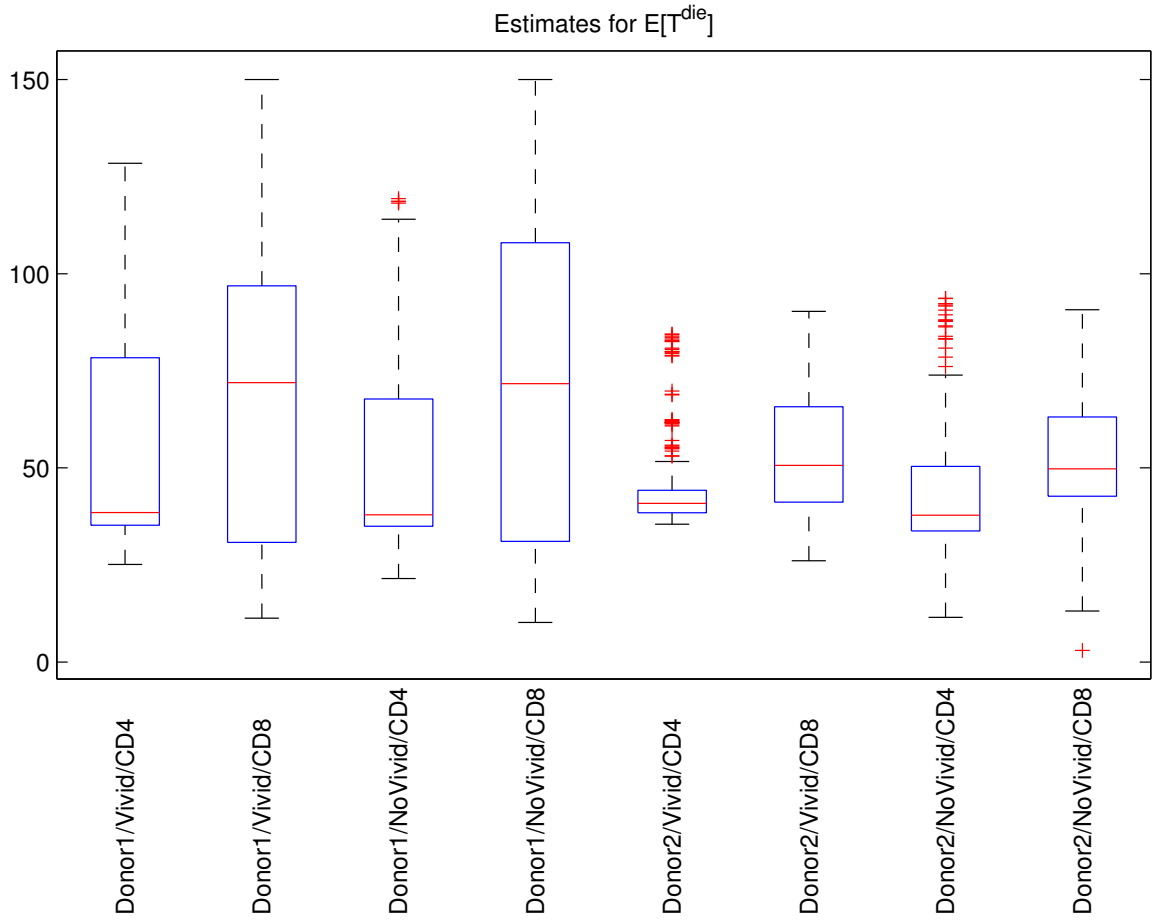


Figure 3.17: Box plots illustrating variability in estimates for the parameter E $\left[T^{die}\right]$.

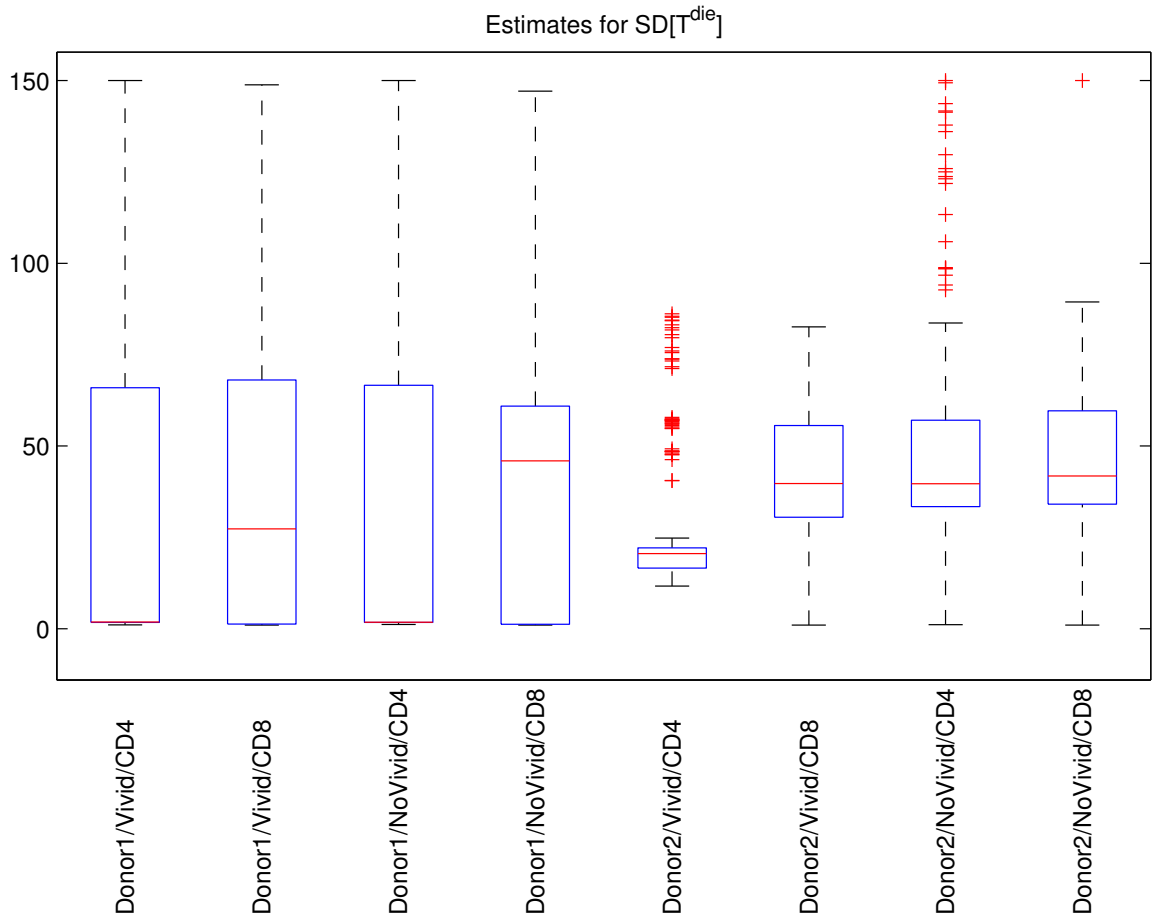


Figure 3.18: Box plots illustrating variability in estimates for the parameter SD $[T^{die}]$.

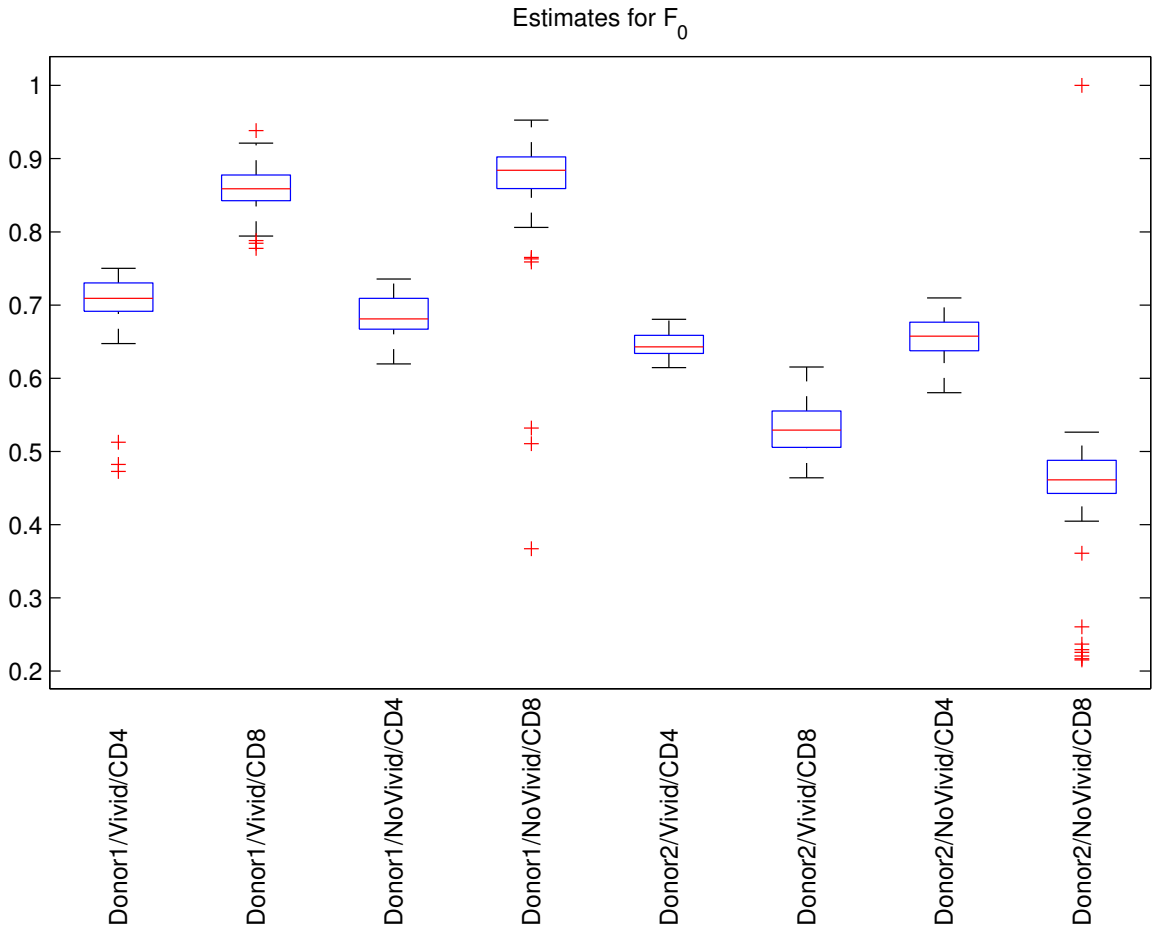


Figure 3.19: Box plots illustrating variability in estimates for the parameter F_0.

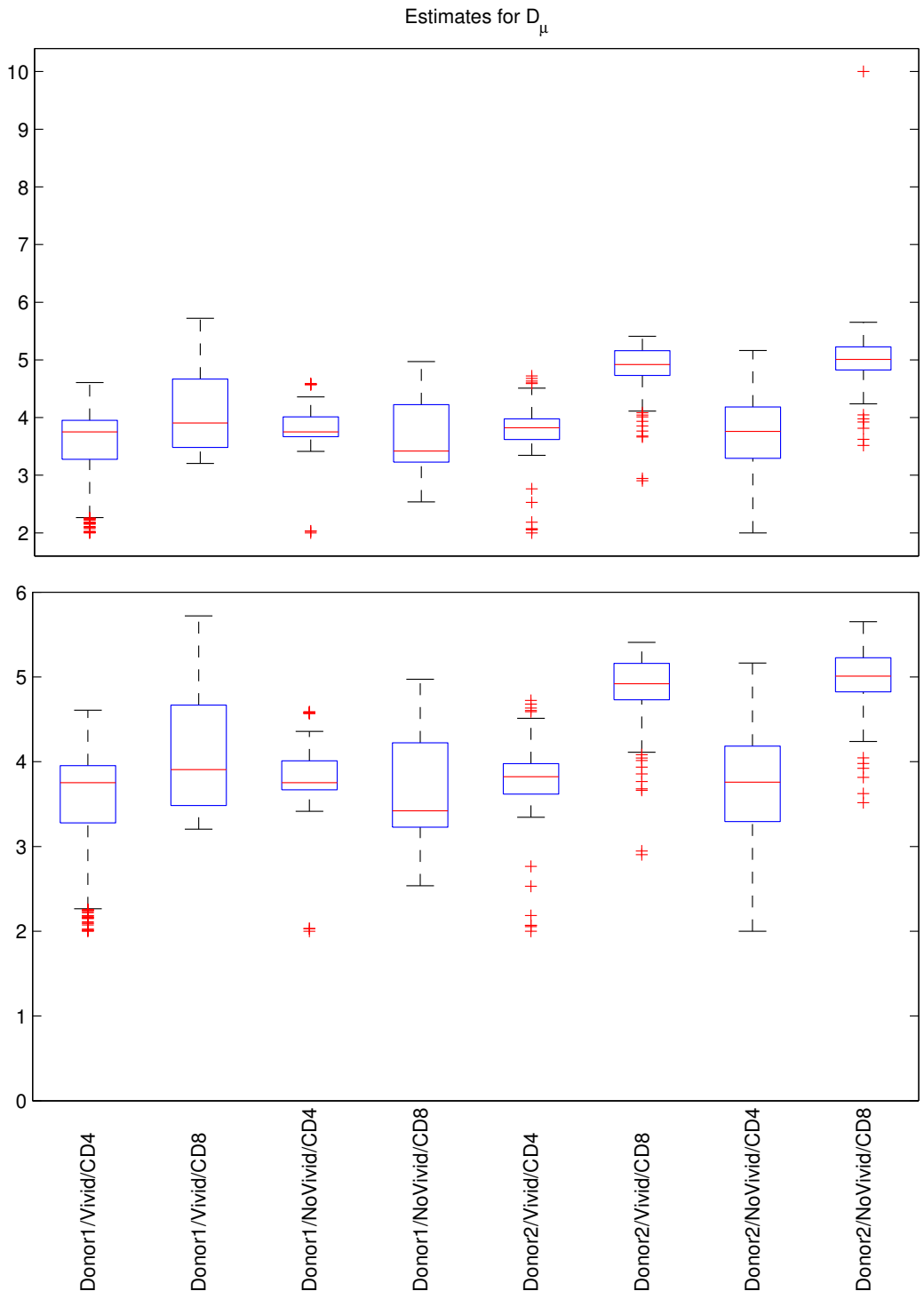


Figure 3.20: Box plots illustrating variability in estimates for the parameter D_μ. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

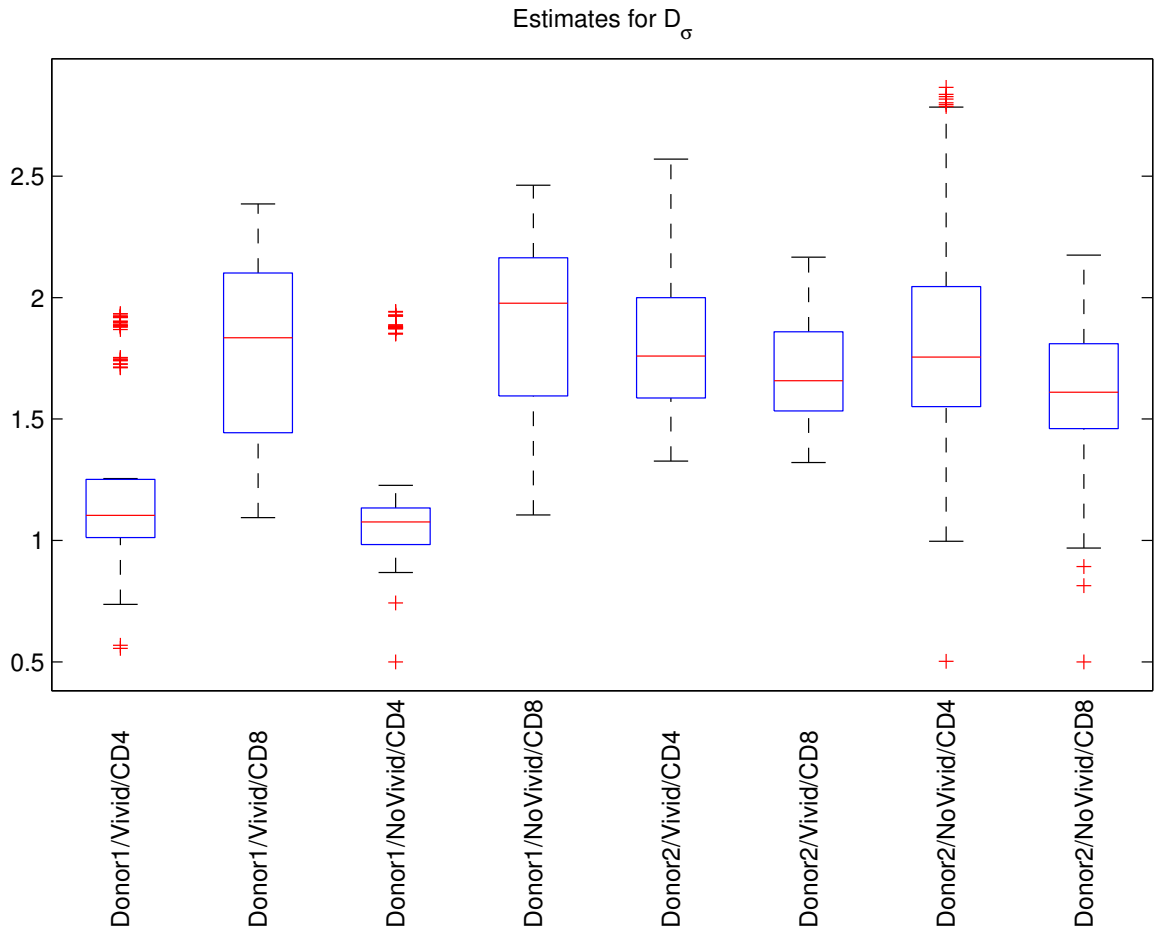


Figure 3.21: Box plots illustrating variability in estimates for the parameter D_σ.

donor and cell type. For example, Figure 3.17 reveals extremely large spreads in many of the box plots corresponding to the parameter $\mathrm{E}\left[T^{die}\right]$, indicating that this model parameter may not be identifiable. On the other hand, all of the box plots corresponding to the parameter c have relatively small spreads, indicating that this model parameter can be estimated with relatively high reliability.

Taken together, all the box plots corresponding to a given parameter allow for useful comparisons of the parameter estimates that are obtained for different combinations of donor, ViViD dye status, and cell type. For example, if we consider the box plots in Figure 3.12 corresponding to the parameter c, we can make a number of interesting conclusions. First, the use of ViViD dye does not appear to lead to a statistically significant difference in the estimate obtained for the parameter c (cf. box plots 1 and 3, 2 and 4, 5 and 7, and 6 and 8, numbering sequentially from left to right). Next, the estimate for c is larger for CD8+ T cells than for CD4+ T cells when considering data for Donor 1 (cf. box plots 1 and 2 or 3 and 4), but it is larger for CD4+ T cells than for CD8+ T cells when considering data for Donor 2 (cf. box plots 5 and 6 or 7 and 8). Finally, while there does not appear to be a statistically significant difference between Donor 1 and Donor 2 in the estimate obtained for c when considering CD4+ T cells (cf. box plots 1 and 5 or 3 and 7), the Donor 1 estimate is considerably larger than the Donor 2 estimate when considering CD8+ T cells (cf. box plots 2 and 6 or 4 and 8).

3.3.1 Remarks on Basic Parameter Estimates

We next expound upon the conclusions that can be drawn by analyzing box plots corresponding to parameter estimates for all 12 of our model parameters. One general conclusion that can be made from these figures is that the use of ViViD dye does not seem to have a large effect on the estimates obtained for most of the model parameters. Therefore, in the discussion that follows we will focus on the box plots summarizing "NoVivid" data sets.

Box plots summarizing estimates for the parameter $\mathrm{E}\left[X_a\right]$, which represents the mean autofluorescence, are shown in Figure 3.10. The 3rd and 4th box plots in that figure indicate that there is a considerable difference in the mean autofluorescence of CD4+ T cells and CD8+ T cells obtained from Donor 1. (Note that the box plots do not "overlap".) More specifically, CD8+ T cells appear to have a larger mean autofluorescence than CD4+ T cells for Donor 1. On the other hand, the 7th and 8th box plots indicate that CD4+ T cells have a larger mean autofluorescence than CD8+ T cells when considering cells obtained from Donor 2. When we compare CD4+ T cells obtained from the two distinct donors (compare 3rd and 7th box plots), it appears that $\mathrm{E}\left[X_a\right]$ is larger for Donor 2 than for Donor 1. When comparing CD8+ T cells from the two donors (compare 4th and 8th box plots), it appears that $\mathrm{E}\left[X_a\right]$ is larger for Donor 1 than for Donor 2.

In Figure 3.11, we provide box plots summarizing estimates for standard deviation of the autofluorescence, SD $[X_a]$. As with the mean autofluorescence, it appears that the value of this parameter is larger for CD8+ T cells than for CD4+ T cells in the case of Donor 1 and larger for CD4+ T cells than for CD8+ T cells in the case of Donor 2. Also, when comparing CD4+ T cells obtained from the two distinct donors, it appears that SD $[X_a]$ is larger for Donor 2 than for Donor 1, and when comparing CD8+ T cells from the two donors, it appears that SD $[X_a]$ is larger for Donor 1 than for Donor 2.

Box plots summarizing estimates for the parameter c, which describes exponential decay of CFSE, are shown in Figure 3.12. It appears that the value of this parameter is larger for CD8+ T cells than for CD4+ T cells in the case of Donor 1 and larger for CD4+ T cells than for CD8+ T cells in the case of Donor 2. When we compare CD4+ T cells obtained from the two distinct donors, there does not appear to be a significant difference in the parameter c; however, when comparing CD8+ T cells from the two donors, it appears that c is larger for Donor 1 than for Donor 2.

In Figure 3.13, we provide box plots summarizing estimates for the parameter E $\left[T_0^{div}\right]$, which represents the mean time to divide for undivided cells. It appears that the value of this parameter is larger for CD4+ T cells than for CD8+ T cells, regardless of which donor we consider. Also, there does not appear to be a significant difference in E $\left[T_0^{div}\right]$ when comparing CD4+ or CD8+ T cells obtained from the two distinct donors. Figure 3.14 indicates that similar statements hold true for SD $\left[T_0^{div}\right]$, which represents the standard deviation in the time to divide for undivided cells.

Box plots summarizing estimates for the parameter E $\left[T^{div}\right]$, which represents the mean time to divide for cells that have divided at least once, are shown in Figure 3.15. There does not appear to be a significant difference in the value of this parameter for CD4+ and CD8+ T cells, whether we consider cells from Donor 1 or Donor 2. On the other hand, when we compare CD4+ or CD8+ T cells from the two distinct donors, it appears that E $\left[T^{div}\right]$ is larger for Donor 1 than for Donor 2.

In Figure 3.16, we provide box plots summarizing estimates for the parameter SD $\left[T^{div}\right]$, which represents the standard deviation in the time to divide for cells that have divided at least once. For this parameter, there does not appear to be a significant difference in the estimated value when comparing the two cell types from a single donor, or when comparing two donors and a single cell type. Considering the widths of the relevant box plots, we see that there is larger variation in the parameter estimates obtained for CD8+ T cells than those obtained for CD4+ T cells. In fact, the variation in parameter estimates observed for Donor 1 CD8+ T cells is so large as to suggest that this parameter is not identifiable.

Box plots summarizing estimates for the parameter E $\left[T^{die}\right]$, which represents the mean time to die for cells that have divided at least once, are shown in Figure 3.17. As was the

case with the parameter SD $\left[T^{div}\right]$, there does not appear to be a significant difference in the estimated value of the parameter E $\left[T^{die}\right]$ when comparing the two cell types from a single donor, or when comparing two donors and a single cell type. For this parameter, we see that there is larger variation in the parameter estimates obtained for Donor 1 cells (of both types) than those obtained for Donor 2 cells. In fact, the variation in parameter estimates observed for Donor 1 cells is so large that it suggests that this parameter is not identifiable.

In Figure 3.18, we provide box plots summarizing estimates for the parameter SD $\left[T^{die}\right]$, which represents the standard deviation in the time to die for cells that have divided at least once. Similar to the situation observed for E $\left[T^{die}\right]$, for this parameter there does not appear to be a significant difference in the estimated value when comparing the two cell types from a single donor, or when comparing two donors and a single cell type. Also similar to the situation observed for E $\left[T^{die}\right]$, we see that there is larger variation in the estimates for SD $\left[T^{die}\right]$ obtained for Donor 1 cells (of both types) than those obtained for Donor 2 cells. The variation in parameter estimates observed for Donor 1 cells is once again so large that it suggests that the parameter SD $\left[T^{die}\right]$ is also not identifiable.

Box plots summarizing estimates for the parameter F_0, which represents the progressor fraction for undivided cells, are shown in Figure 3.19. It appears that the value of this parameter is larger for CD8+ T cells than for CD4+ T cells in the case of Donor 1 and larger for CD4+ T cells than for CD8+ T cells in the case of Donor 2. When we compare CD4+ T cells obtained from the two distinct donors, there does not appear to be a significant difference in the parameter F_0; however, when comparing CD8+ T cells from the two donors, it appears that F_0 is larger for Donor 1 than for Donor 2.

In Figure 3.20, we provide box plots summarizing estimates for the parameter D_μ. There does not appear to be a significant difference in the value of this parameter for CD4+ and CD8+ T cells in the case of Donor 1, but the parameter value is larger for CD8+ T cells in the case of Donor 2. When we compare CD4+ T cells obtained from the two distinct donors, there does not appear to be a significant difference in the parameter D_μ; however, when comparing CD8+ T cells from the two donors, it appears that D_μ is larger for Donor 2 than for Donor 1.

Box plots summarizing estimates for the parameter D_σ are shown in Figure 3.21. It appears that the parameter value is larger for CD8+ T cells than CD4+ T cells in the case of Donor 1, but there does not appear to be a significant difference in the parameter value for CD4+ and CD8+ T cells in the case of Donor 2. When we compare CD4+ T cells obtained from the two distinct donors, it appears that D_σ is larger for Donor 2 than for Donor 1. On the other hand, when comparing CD8+ T cells from the two donors, there does not appear to be a significant difference in the parameter D_σ.

We have already noted that many of our model parameters can be estimated with relatively high reliability, while others do not appear to be identifiable, but thus far all of our arguments

have been based upon visual inspection of the box plots in Figures 3.10 through 3.21. To allow for more careful quantitative analysis of the identifiability of parameters, we provide in Tables 3.6 through 3.17 some of the summary statistics used to generate the box plots in those figures. For example, in Table 3.6 we provide a median and an interquartile range (IQR) corresponding to each of the box plots in Figure 3.10. Since the median and IQR indicate the "center" and "spread", respectively, for a set of parameter estimates, the ratio of these two quantities provides a useful measure of "relative spread". We therefore also include a column for the ratio of IQR to median in each of Tables 3.6 through 3.17. When the spread is greater than 50% of the central value for a particular set of parameter estimates (i.e., whenever the ratio of IQR to median is greater than 0.50), we consider the variability in that set of parameter estimates to be "relatively high" and conclude that the parameter may not be identifiable; therefore, the ratios meeting this criteria are emphasized in boldface in the tables. Note that the value 0.50 was chosen somewhat arbitrarily, but comparing Tables 3.6 through 3.17 with Figures 3.10 through 3.21 makes it clear that such a value for the ratio of IQR to median does, indeed, indicate a "large" relative spread in a set of parameter estimates.

Based on Tables 3.6 through 3.17, we conclude that the model parameters $\mathrm{E}[X_a]$, $\mathrm{SD}[X_a]$, c, $\mathrm{E}\left[T_0^{div}\right]$, $\mathrm{SD}\left[T_0^{div}\right]$, $\mathrm{E}\left[T^{div}\right]$, F_0, D_μ, and D_σ can all be estimated with fairly high reliability. On the other hand, the parameters $\mathrm{SD}\left[T^{div}\right]$, $\mathrm{E}\left[T^{die}\right]$, and $\mathrm{SD}\left[T^{die}\right]$ each have very high ratios of IQR to median in some cases, indicating that they may not be identifiable. We conjecture that this could be because the mathematical model is not sensitive to these particular parameters. One might reason, for example, that the lack of model sensitivity to the parameters involving "time until death" occurs because divided cells (those cells for which $i \geq 1$) tend to divide much more often than they die (when considering stimulated T cells from healthy donors) and such behavior can be correctly incorporated into the model as long as the expected time until division, $\mathrm{E}\left[T^{div}\right]$, is significantly smaller than the expected time until death, $\mathrm{E}\left[T^{die}\right]$. To be more specific (and to reuse some terminology that was employed in Section 3.2), as long as the distribution of the random variable T^{die} tends to be *stochastically larger* (by a substantial margin) than that of T^{div}, the correct dynamical behavior of the system can probably be modeled adequately (at least over the first few days) *even if the parameters describing the distribution of* T^{die} *are not estimated very accurately.* To illustrate this point, consider Figures 3.22 and 3.23, in which we plot the (lognormal) distributions of the random variables T^{div} and T^{die} in two different cases. In both cases we use the same set of parameter values for T^{div}, but in each case a different set of a parameter values is used for T^{die}. The values of T^{die} tend to be larger than those of T^{div} in both cases, so in both situations cells should tend to divide more frequently than they die. (Recall that the fate of any particular cell is determined by whichever of these two random variables produces a *smaller* realization.) We argue that, when all other parameters are fixed using some common set of values, the model output does not vary significantly (at

Table 3.6: Summary statistics for estimates of parameter $\mathrm{E}[X_a]$ (cf. Figure 3.10).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	402.30	32.58	0.0810
1	Y	CD8	680.44	79.29	0.1165
1	N	CD4	388.45	25.33	0.0652
1	N	CD8	678.97	81.78	0.1204
2	Y	CD4	623.98	50.55	0.0810
2	Y	CD8	542.52	21.49	0.0396
2	N	CD4	630.26	82.19	0.1304
2	N	CD8	536.94	24.43	0.0455

Table 3.7: Summary statistics for estimates of parameter $\mathrm{SD}[X_a]$ (cf. Figure 3.11).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	227.87	17.03	0.0747
1	Y	CD8	339.49	78.22	0.2304
1	N	CD4	215.83	25.02	0.1159
1	N	CD8	335.50	78.12	0.2328
2	Y	CD4	293.16	22.61	0.0771
2	Y	CD8	252.75	15.51	0.0614
2	N	CD4	294.12	39.91	0.1357
2	N	CD8	253.24	16.62	0.0656

Table 3.8: Summary statistics for estimates of parameter c (cf. Figure 3.12).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	0.0069	0.0003	0.0461
1	Y	CD8	0.0077	0.0004	0.0520
1	N	CD4	0.0067	0.0003	0.0487
1	N	CD8	0.0073	0.0004	0.0609
2	Y	CD4	0.0068	0.0003	0.0448
2	Y	CD8	0.0039	0.0006	0.1424
2	N	CD4	0.0065	0.0002	0.0343
2	N	CD8	0.0040	0.0004	0.0932

Table 3.9: Summary statistics for estimates of parameter E $\left[T_0^{div}\right]$ (cf. Figure 3.13).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	63.02	2.28	0.0362
1	Y	CD8	50.20	2.34	0.0466
1	N	CD4	62.10	1.92	0.0308
1	N	CD8	49.68	3.61	0.0726
2	Y	CD4	60.26	1.36	0.0226
2	Y	CD8	48.49	1.06	0.0218
2	N	CD4	63.42	2.23	0.0351
2	N	CD8	47.54	1.07	0.0226

Table 3.10: Summary statistics for estimates of parameter SD $\left[T_0^{div}\right]$ (cf. Figure 3.14).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	28.07	1.63	0.0580
1	Y	CD8	19.21	1.92	0.0998
1	N	CD4	27.20	1.79	0.0658
1	N	CD8	18.08	4.30	0.2377
2	Y	CD4	23.90	1.00	0.0418
2	Y	CD8	18.56	0.89	0.0478
2	N	CD4	26.62	1.91	0.0719
2	N	CD8	17.49	0.81	0.0464

Table 3.11: Summary statistics for estimates of parameter E $\left[T^{div}\right]$ (cf. Figure 3.15).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	10.96	0.74	0.0672
1	Y	CD8	10.39	1.33	0.1276
1	N	CD4	11.21	0.70	0.0623
1	N	CD8	10.89	1.79	0.1640
2	Y	CD4	8.96	0.81	0.0908
2	Y	CD8	8.97	0.51	0.0569
2	N	CD4	8.97	0.75	0.0833
2	N	CD8	9.20	0.46	0.0505

Table 3.12: Summary statistics for estimates of parameter SD $\left[T^{div}\right]$ (cf. Figure 3.16).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	0.86	0.73	**0.8428**
1	Y	CD8	1.47	4.09	**2.7912**
1	N	CD4	0.89	0.12	0.1342
1	N	CD8	4.25	5.67	**1.3347**
2	Y	CD4	0.80	0.08	0.0988
2	Y	CD8	0.94	0.44	0.4648
2	N	CD4	0.82	0.04	0.0507
2	N	CD8	1.00	0.61	**0.6087**

Table 3.13: Summary statistics for estimates of parameter E $\left[T^{die}\right]$ (cf. Figure 3.17).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	38.49	43.14	**1.1207**
1	Y	CD8	71.96	66.02	**0.9175**
1	N	CD4	37.90	32.78	**0.8650**
1	N	CD8	71.68	76.91	**1.0729**
2	Y	CD4	40.84	5.82	0.1426
2	Y	CD8	50.67	24.57	0.4848
2	N	CD4	37.83	16.64	0.4397
2	N	CD8	49.75	20.40	0.4100

Table 3.14: Summary statistics for estimates of parameter SD $\left[T^{die}\right]$ (cf. Figure 3.18).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	1.85	64.21	**34.7682**
1	Y	CD8	27.30	66.82	**2.4481**
1	N	CD4	1.79	64.90	**36.1615**
1	N	CD8	45.91	59.70	**1.3004**
2	Y	CD4	20.56	5.54	0.2693
2	Y	CD8	39.72	25.11	**0.6322**
2	N	CD4	39.68	23.65	**0.5960**
2	N	CD8	41.82	25.57	**0.6113**

Table 3.15: Summary statistics for estimates of parameter F_0 (cf. Figure 3.19).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	0.7092	0.0388	0.0547
1	Y	CD8	0.8588	0.0350	0.0408
1	N	CD4	0.6812	0.0421	0.0619
1	N	CD8	0.8840	0.0434	0.0491
2	Y	CD4	0.6429	0.0248	0.0386
2	Y	CD8	0.5293	0.0496	0.0937
2	N	CD4	0.6574	0.0392	0.0597
2	N	CD8	0.4613	0.0453	0.0982

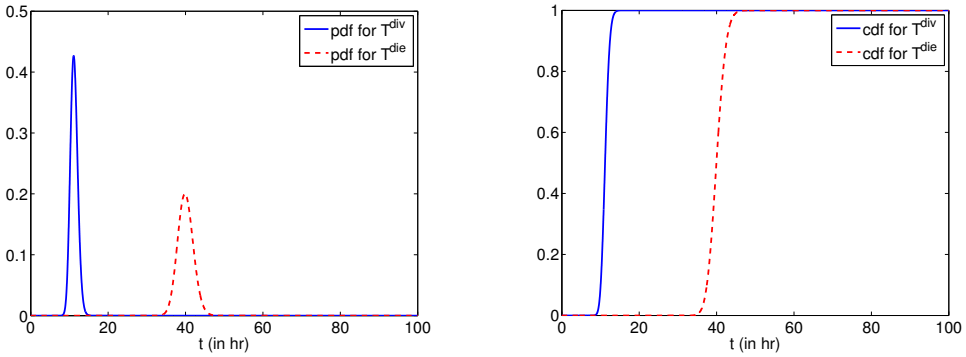
Table 3.16: Summary statistics for estimates of parameter D_μ (cf. Figure 3.20).

Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	3.75	0.68	0.1801
1	Y	CD8	3.91	1.19	0.3034
1	N	CD4	3.75	0.34	0.0913
1	N	CD8	3.42	1.00	0.2913
2	Y	CD4	3.82	0.36	0.0939
2	Y	CD8	4.92	0.43	0.0874
2	N	CD4	3.76	0.89	0.2373
2	N	CD8	5.01	0.40	0.0803

Table 3.17: Summary statistics for estimates of parameter D_σ (cf. Figure 3.21).

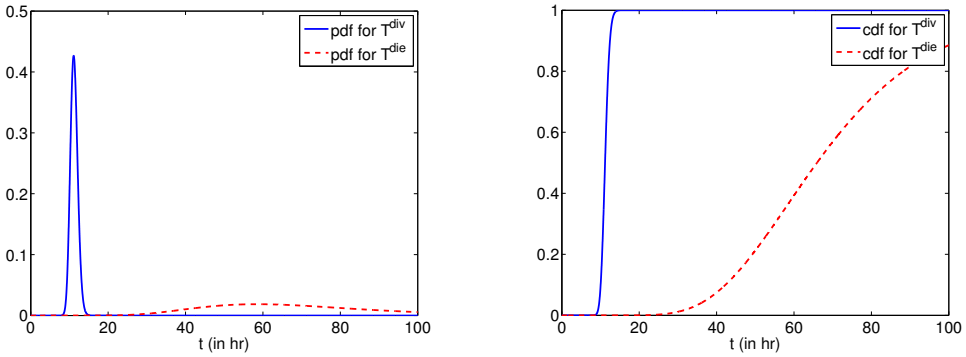
Donor	ViViD Used	Cell Type	Median	IQR	IQR/Median
1	Y	CD4	1.10	0.24	0.2172
1	Y	CD8	1.83	0.66	0.3583
1	N	CD4	1.08	0.15	0.1398
1	N	CD8	1.98	0.57	0.2877
2	Y	CD4	1.76	0.41	0.2349
2	Y	CD8	1.66	0.33	0.1968
2	N	CD4	1.76	0.49	0.2813
2	N	CD8	1.61	0.35	0.2169

least over the first few days) when the two different sets of parameter values for T^{die} indicated in Figures 3.22 and 3.23 are used. To demonstrate this claim, we show sample model output generated using these two different sets of parameter values for T^{die} in Figure 3.24. The complete sets of parameter values for "Model A" and "Model B" are provided in Table 3.18. Note that, despite the large discrepancy in the values used for E $\left[T^{die}\right]$ and SD $\left[T^{die}\right]$, the output for the two models is indistinguishable until at least Day 3. This simple exercise indicates that the accuracy of data collected in the later days of the experiment may be critical to the correct identification of the parameters E $\left[T^{die}\right]$ and SD $\left[T^{die}\right]$. In Section 3.3.2, we test this hypothesis using model comparison tests.



(a) (b)

Figure 3.22: Plots illustrating the (a) pdfs and (b) cdfs of the lognormally distributed random variables T^{div} and T^{die} when E $\left[T^{div}\right] = 11.21$, SD $\left[T^{div}\right] = 0.89$, E $\left[T^{die}\right] = 40$, and SD $\left[T^{die}\right] = 2$.



(a) (b)

Figure 3.23: Plots illustrating the (a) pdfs and (b) cdfs of the lognormally distributed random variables T^{div} and T^{die} when E $\left[T^{div}\right] = 11.21$, SD $\left[T^{div}\right] = 0.89$, E $\left[T^{die}\right] = 70$, and SD $\left[T^{die}\right] = 25$.

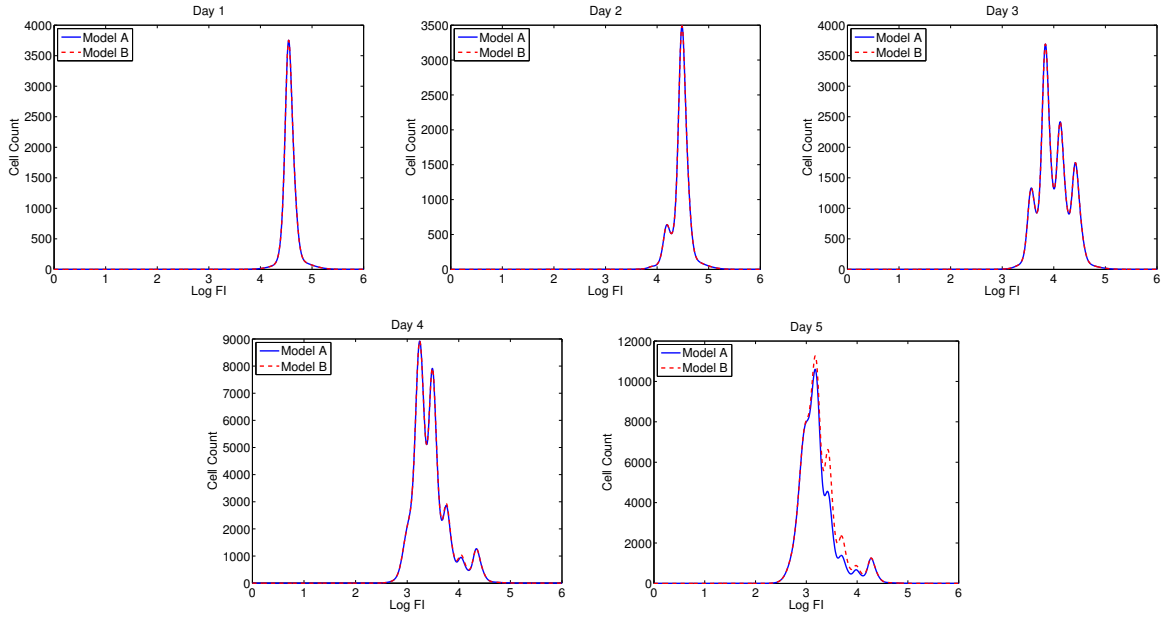


Figure 3.24: Summary histogram output for Model A, in which $\mathrm{E}\left[T^{die}\right] = 40$ and $\mathrm{SD}\left[T^{die}\right] = 2$, and Model B, in which $\mathrm{E}\left[T^{die}\right] = 70$ and $\mathrm{SD}\left[T^{die}\right] = 25$. The time points corresponding to "Day 1" through "Day 5" are the same as those that were used for data collection as described in Section 3.1.

Table 3.18: Parameter values used to obtain summary histogram output for Models A and B in Figure 3.24. Values which differ in the two models are emphasized in boldface.

Parameter	Model A	Model B
$\mathrm{E}\left[X_a\right]$	388.45	388.45
$\mathrm{SD}\left[X_a\right]$	215.83	215.83
c	0.0067	0.0067
$\mathrm{E}\left[T_0^{div}\right]$	62.10	62.10
$\mathrm{SD}\left[T_0^{div}\right]$	27.20	27.20
$\mathrm{E}\left[T^{div}\right]$	11.21	11.21
$\mathrm{SD}\left[T^{div}\right]$	0.89	0.89
$\mathrm{E}\left[T^{die}\right]$	**40.00**	**70.00**
$\mathrm{SD}\left[T^{die}\right]$	**2.00**	**25.00**
F_0	0.6812	0.6812
D_μ	3.75	3.75
D_σ	1.08	1.08

3.3.2 Qualifying Identifiability of T^{die} Parameters Using Model Comparison Tests

In this section, we first seek to demonstrate that the parameters $\mathrm{E}\left[T^{die}\right]$ and $\mathrm{SD}\left[T^{die}\right]$ *are* influential in describing the behavior of a population of proliferating cells over five days of an experiment such as the one described in Section 3.1. We then attempt to show that the same parameters are *not* influential (under certain conditions) in describing behavior during the first three days of such experiments. The example provided at the end of Section 3.3.1 attests to the plausibility of these hypotheses by showing that model output for the first three days is not significantly affected when the parameters in question are changed, while the output for Day 5 does appear to be sensitive to changes in these parameters.

In order to test the hypothesis that the parameters $\mathrm{E}\left[T^{die}\right]$ and $\mathrm{SD}\left[T^{die}\right]$ are non-influential, we can use statistically based model comparison techniques. The approach we use is based on analysis of variance (ANOVA) hypothesis testing as outlined by Banks and Tran in their text on mathematical modeling [12]. We consider two distinct mathematical models, both of which can be evaluated using the cost functional J given in (2.17). The first is the twelve-parameter model described in Section 2.5 and the second is the "nested" model that results when the parameters $\mathrm{E}\left[T^{die}\right]$ and $\mathrm{SD}\left[T^{die}\right]$ are fixed at 70 hours and 1 hour, respectively. Note that these fixed parameter values tend to ensure that T^{die} will be stochastically larger than T^{div} when calibrating the nested model using the data described in Section 3.1.

In order to formulate our statistical hypotheses, we let $\mathcal{Q} \subset \mathbb{R}^{12}$ denote the set of all admissible parameters for the twelve-parameter model and $\mathcal{Q}^H = \{\vec{q} \in \mathcal{Q} : H\vec{q} = \vec{c}\} \subset \mathcal{Q}$ be the set of admissible parameters for the nested model, where $H \in \mathbb{R}^{2\times 12}$ and $\vec{c} \in \mathbb{R}^2$. Using the parameter ordering suggested by Table 2.1 and setting

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

and $\vec{c} = (70, 1)^T$ results in the nested model described previously. We wish to test the null hypothesis that the "true" parameter vector $\vec{q}_0$ is in the restricted set $\mathcal{Q}^H$; i.e.,

$$H_0 : \vec{q}_0 \in \mathcal{Q}^H.$$

Therefore, we follow the usual practice in inferential statistics of defining a test statistic. Let $\{N_k^j\}$ be a set of random variables as in (2.16) with corresponding realizations $\{n_k^j\}$ constituting observed data so that we can define the GLS estimators

$$\vec{q}_{GLS} = \underset{\vec{q}\in\mathcal{Q}}{\operatorname{argmin}}\, J(\vec{q}; \{N_k^j\}) \quad \text{and} \quad \vec{q}_{GLS}^{\,H} = \underset{\vec{q}\in\mathcal{Q}^H}{\operatorname{argmin}}\, J(\vec{q}; \{N_k^j\}) \tag{3.1}$$

and GLS estimates

$$\hat{q}_{GLS} = \underset{\vec{q}\in\mathcal{Q}}{\operatorname{argmin}}\, J(\vec{q};\{n_k^j\}) \quad \text{and} \quad \hat{q}_{GLS}^H = \underset{\vec{q}\in\mathcal{Q}^H}{\operatorname{argmin}}\, J(\vec{q};\{n_k^j\}). \tag{3.2}$$

We note here that $J(\hat{q}_{GLS}^H;\{n_k^j\}) \geq J(\hat{q}_{GLS};\{n_k^j\})$ because the estimate $\hat{q}_{GLS}^H$ is obtained by optimizing over a subset of $\mathcal{Q}$, while $\hat{q}_{GLS}$ is obtained by optimizing over all of $\mathcal{Q}$. Using the GLS estimators and estimates, we can define the test statistic

$$U(\{N_k^j\}) = \frac{n\Big(J(\vec{q}_{GLS}^H,\{N_k^j\}) - J(\vec{q}_{GLS},\{N_k^j\})\Big)}{J(\vec{q}_{GLS},\{N_k^j\})} \tag{3.3}$$

with corresponding realization

$$\hat{U}(\{n_k^j\}) = \frac{n\Big(J(\hat{q}_{GLS}^H,\{n_k^j\}) - J(\hat{q}_{GLS},\{n_k^j\})\Big)}{J(\hat{q}_{GLS},\{n_k^j\})}, \tag{3.4}$$

where

$$n = [\text{number of histogram bins}] \times [\text{number of time points}]$$

is the number of observations in a data set. We remark that, for the five-day time series data sets considered in this study, we use 1024 bins and $5 - 1 = 4$ time points (because the first time point is used to construct an initial condition and is not considered to be a data point), so $n = 1024 \times 4 = 5096$. As discussed by Banks and Tran [12], the test statistic U converges in distribution to a χ^2 distribution with $r = 2$ degrees of freedom (where r is the number of constraints defined by the system $H\vec{q} = \vec{c}$) as $n \to \infty$.

As shown in Table 3.19, the costs associated with the optimal parameter vector $\hat{q}_{GLS}^H$ on the restricted set $\mathcal{Q}^H$ tend to be significantly greater than those associated with with the optimal parameter vector $\hat{q}_{GLS}$ on the set $\mathcal{Q}$ when all data points up through Day 5 are considered. The results shown in the table were obtained for data corresponding to Donor 1 CD4+ T cells when ViViD dye is not used to exclude dead cells (i.e., experimental condition "Donor1/NoVivid/CD4"), but they are typical of the results obtained with five-day time series data sets for any of the eight combinations of donor, ViViD dye status, and cell type. We also remark that Table 3.19 only shows results for 27 of the 162 possible five-day data sets that can be formed for experimental condition Donor1/NoVivid/CD4. The "Data ID" column in the table provides a 5-digit base 3 number that indicates which triplicate sample ("0" for Sample 1, "1" for Sample 2, or "2" for Sample 3) is used for each of the five days. Again, results for the $162 - 27 = 135$ five-day data sets *not* shown in the table are similar to those that are shown. Based on the very low (essentially zero) p-values resulting from the model comparison tests, we

Table 3.19: Results of the model comparison test described in Section 3.3.2 when using all data points up through Day 5.

Data ID	$J(\hat{q}_{GLS}, \{n_k^j\})$	$J(\hat{q}_{GLS}^H, \{n_k^j\})$	$\hat{U}(\{n_k^j\})$	p-value
00000	19875.8	28477.6	1772.65	0
00001	19070.2	23057.3	856.383	0
00002	19405	24010.1	972.061	0
00010	22566	31208.3	1568.68	0
00011	16521.6	18404.3	466.774	0
00012	20302.2	28115.2	1576.28	0
00100	22971.1	29463.4	1157.63	0
00101	20273.5	24604.9	875.102	0
00102	20473.2	25452	996.076	0
00110	23429.4	32159.7	1526.26	0
00111	17469.9	19466.3	468.082	0
00112	21247.3	29062.9	1506.68	0
00200	22333.3	28423.7	1117.01	0
00201	20574.8	24777.8	836.73	0
00202	20435	24502.9	815.374	0
00210	23602.9	31881.7	1436.69	0
00211	17485.5	19339.4	434.279	0
00212	21403.2	28846.5	1424.44	0
01000	22542.9	29237.8	1216.45	0
01001	20365	24334.7	798.427	0
01002	20807.2	25238.5	872.326	0
01010	23732.8	32154.3	1453.45	0
01011	17762.9	19641.3	433.153	0
01012	17796.4	20048.2	518.28	0
01100	24356.4	30714.8	1069.29	0
01101	21054.5	24710.3	711.226	0
01102	21660.7	26724.7	957.593	0

reject the null hypothesis and infer that the T^{die} parameters *are* important for describing the behavior of a population of proliferating cells over five days.

The hypothesis test results are quite different when we only consider data from the first three days of cell proliferation. As shown in Table 3.20, the costs associated with the optimal

parameter vector $\hat{q}^H_{GLS}$ on the restricted set $\mathcal{Q}^H$ tend to be very close to those associated with with the optimal parameter vector $\hat{q}_{GLS}$ on the set $\mathcal{Q}$ when only data points for Days 1 through 3 are considered. We remark that in some cases $J(\hat{q}^H_{GLS}, \{n^j_k\}) < J(\hat{q}_{GLS}, \{n^j_k\})$ leading to a negative value for the test statistic $\hat{U}(\{n^j_k\})$. While this should *theoretically* not occur, we point out that all estimates are obtained through numerical optimization (cf. Section 2.6). The differences in the two costs in these cases is usually quite small and is therefore probably attributable to tolerances used in the numerical parameter estimation routines. Again, the results shown in the table were obtained for experimental condition Donor1/NoVivid/CD4, but they are typical of the results obtained with three-day time series data sets for any of the eight combinations of donor, ViViD dye status, and cell type. In some extreme cases (not shown in the table) $J(\hat{q}_{GLS}, \{n^j_k\})$ is significantly smaller that $J(\hat{q}^H_{GLS}, \{n^j_k\})$, but this could be explained by other numerical issues; for example, the optimization algorithm involved in parameter estimation might fail to find a global minimum if it first arrives at a local minimum associated with the fixed parameter values. Despite these numerical issues, the result in the majority of the model comparison tests is a very high p-value. Thus, we fail to reject the null hypothesis and infer that that the T^{die} parameters are *not* important for describing the behavior of a population of proliferating cells during Days 1 through 3.

The outcome of the model comparison tests described here can be summarized as follows. While the T^{die} parameters do not appear to significantly impact the behavior of a population of proliferating cells during Days 1 through 3, they do appear to be important in describing the status of the population at later time points. Thus, if reliable (replicable) data could be obtained for the later time points (Days 4 and 5), the T^{die} parameters could probably be identified.

Table 3.20: Results of the model comparison test described in Section 3.3.2 when using data points for Days 1 through 3.

Data ID	$J(\hat{q}_{GLS}, \{n_k^j\})$	$J(\hat{q}_{GLS}^H, \{n_k^j\})$	$\hat{U}(\{n_k^j\})$	p-value
000	7658.73	7657.05	-0.448661	1
001	8351.69	8348.66	-0.741913	1
002	8147.56	8147.74	0.0462316	0.977149
010	8693.52	8693.04	-0.112691	1
011	9383.89	9386.68	0.607774	0.737944
012	9331.61	9179.68	-33.3447	1
020	10679.8	10680.6	0.150055	0.927718
021	11408.3	11409.4	0.182332	0.912866
022	11206.4	11203.7	-0.484748	1
100	5572.97	5574.21	0.456138	0.796069
101	6036.02	6037.61	0.541507	0.762804
102	5887.39	5887.25	-0.0483848	1
110	6008.93	6009.26	0.115165	0.944044
111	6467.57	6467.93	0.112524	0.945292
112	6321.62	6321	-0.200943	1
120	7416.37	7417.08	0.197622	0.905914
121	7906.26	7906.04	-0.0570808	1
122	7763.63	7763.01	-0.165685	1
200	4912.94	4913.64	0.29154	0.864356
201	5288.85	5287.78	-0.415375	1
202	5152.64	5152.67	0.00871946	0.99565
210	5107.48	5107.84	0.144135	0.930468
211	5480.76	5481.77	0.377473	0.828004
212	5345.85	5345.61	-0.0897012	1
220	6062.62	6062.55	-0.0234248	1
221	6458.76	6458.47	-0.0919566	1
222	6327.65	6328.04	0.125123	0.939355

3.3.3 Parameter Estimates Obtained Using Fixed Values for Some Parameters

By examining scatter plots of various pairings of parameter estimates, we can determine whether or not any correlations might exist between some of the parameters. For example, Figures 3.25 and 3.26 indicate a strong correlation between the parameters $\mathrm{E}\left[T_0^{div}\right]$ and $\mathrm{SD}\left[T_0^{div}\right]$ and $\mathrm{E}\left[T^{die}\right]$ and $\mathrm{SD}\left[T^{die}\right]$, respectively. We therefore use once again a variation of our parameter estimation scheme in which one or more of the 12 model parameters can be fixed (cf. Section 3.3.2), hoping that the fixing of certain parameters might reduce the variability seen in some of the other parameter estimates.

Based on the relationships suggested by Figures 3.25 and 3.26, we applied our modified parameter estimation algorithm in two scenarios: (i) using fixed values for $\mathrm{E}\left[T_0^{div}\right]$, and (ii) using fixed values for $\mathrm{E}\left[T^{die}\right]$. We chose to fix the values of the two selected parameters at the (approximate) median estimates obtained from the basic parameter estimation scheme (with no fixed parameters). Since these medians vary for the different combinations of donor, ViViD dye status, and cell type, we used different fixed values for each of these combinations. The specific fixed values used for $\mathrm{E}\left[T_0^{div}\right]$ and $\mathrm{E}\left[T^{die}\right]$ are shown in Tables 3.21 and 3.22, respectively. Our goal in fixing the value of one or more parameters is to reduce the variability in the parameter estimates for other parameters, so we need to select some measure of variability for the analysis. Throughout the discussions that follow, we will use the interquartile range (IQR) as a rough measure of variability in the parameter estimates for a given combination of donor, ViViD dye status, and cell type.

Table 3.21: Parameter values used when fixing the parameter $\mathrm{E}\left[T_0^{div}\right]$.

Donor	ViViD Used	Cell Type	$\mathrm{E}\left[T_0^{div}\right]$
1	Y	CD4	64
1	Y	CD8	51
1	N	CD4	62.5
1	N	CD8	49
2	Y	CD4	60.5
2	Y	CD8	48
2	N	CD4	64
2	N	CD8	47

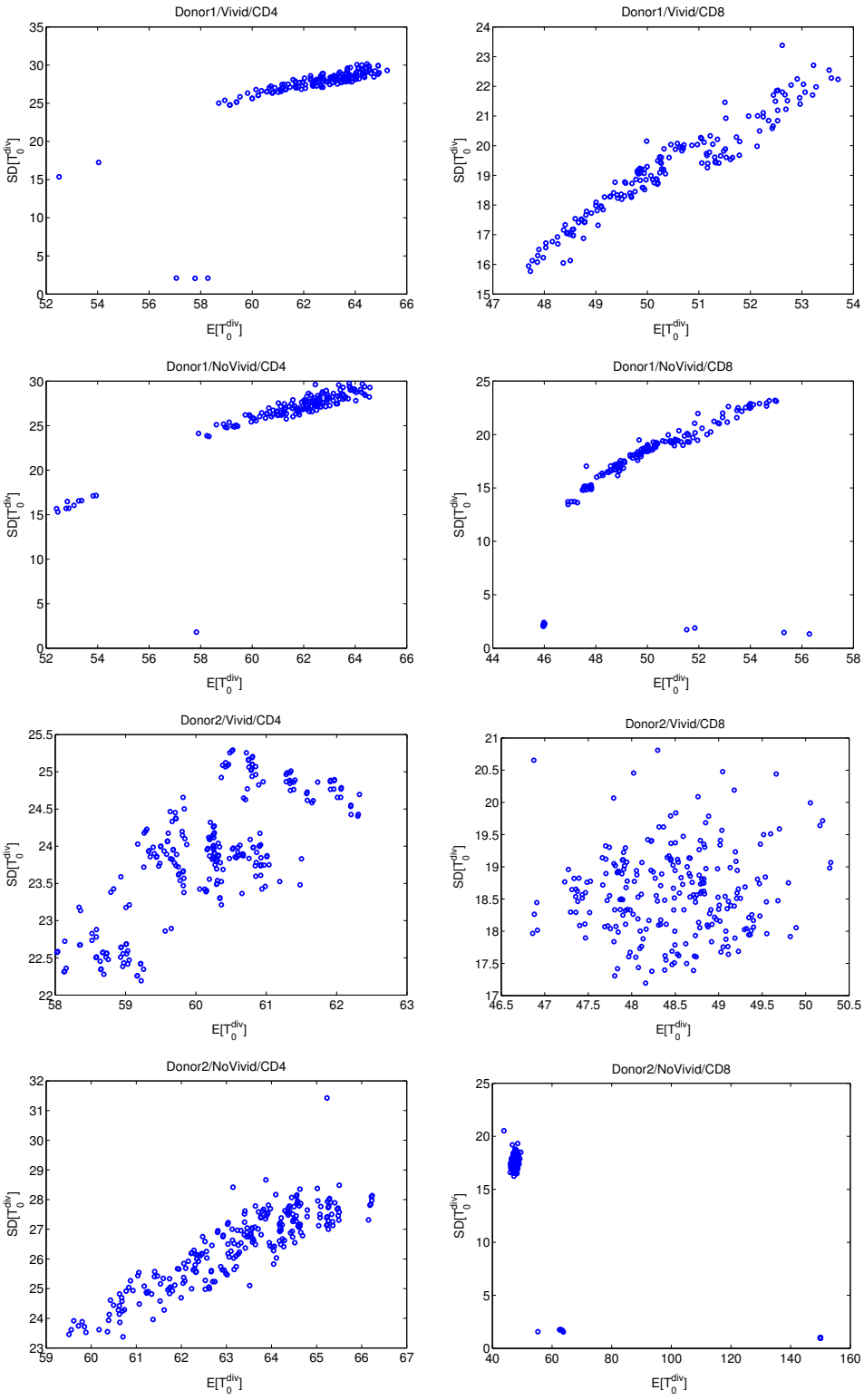


Figure 3.25: Scatterplots illustrating a correlation between E $\left[T_0^{div}\right]$ and SD $\left[T_0^{div}\right]$.

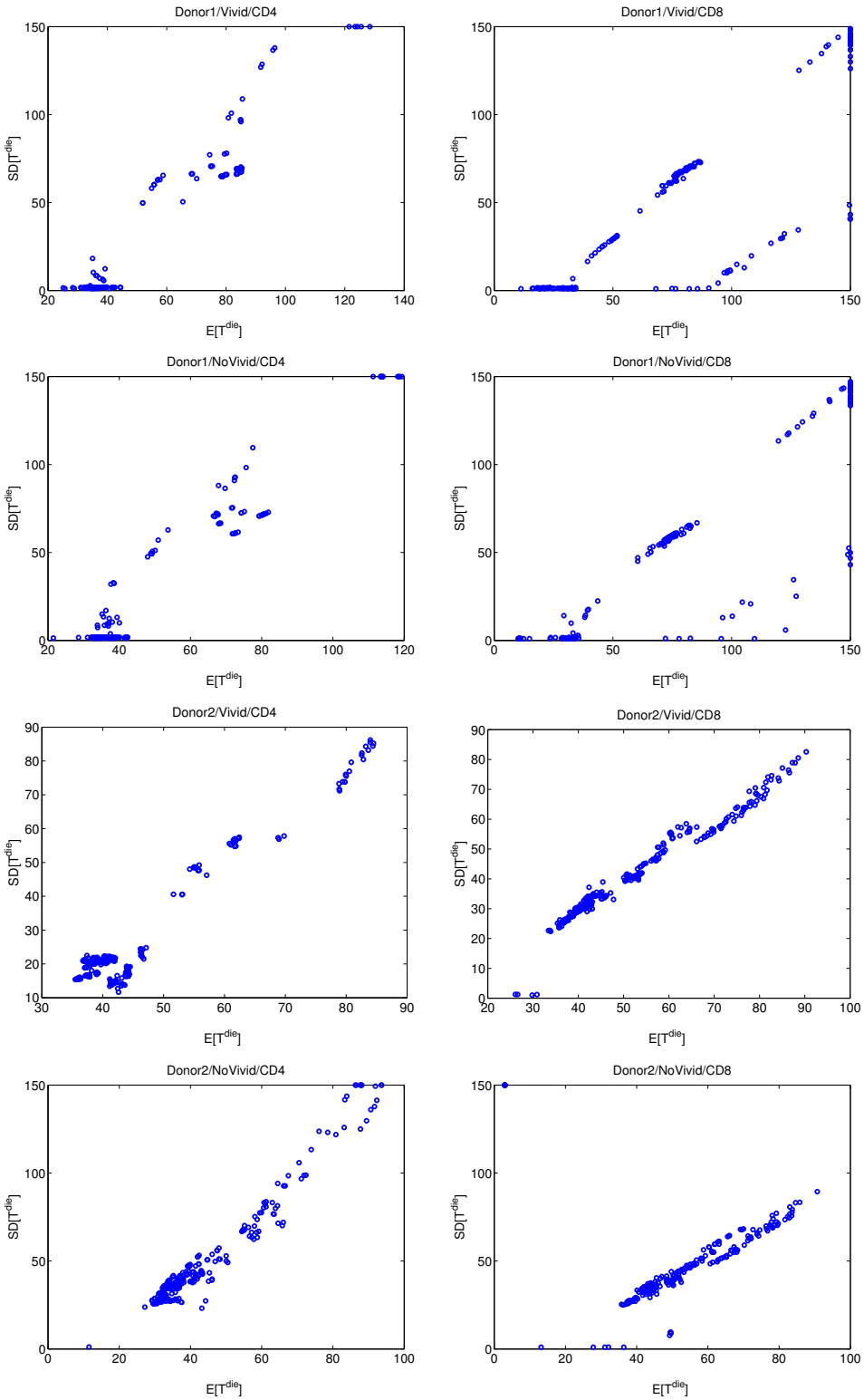


Figure 3.26: Scatterplots illustrating a correlation between E $\left[T^{die}\right]$ and SD $\left[T^{die}\right]$.

Table 3.22: Parameter values used when fixing the parameter $\mathrm{E}\left[T^{die}\right]$.

Donor	ViViD Used	Cell Type	$\mathrm{E}\left[T^{die}\right]$
1	Y	CD4	40
1	Y	CD8	62
1	N	CD4	40
1	N	CD8	62
2	Y	CD4	40
2	Y	CD8	52
2	N	CD4	36
2	N	CD8	50

The results of applying our modified parameter estimation technique in case (i) are shown in Figures 3.27 through 3.37. Note that the total number of figures is 11 – there is one for each model parameter except $\mathrm{E}\left[T_0^{div}\right]$, which is fixed. Comparing Figures 3.14 and 3.30, we see that using a fixed value for $\mathrm{E}\left[T_0^{div}\right]$ does considerably reduce the variability in the estimates of $\mathrm{SD}\left[T_0^{div}\right]$ in the case of Donor 1 data; however, this is not generally true in the case of Donor 2 data. In fact, when using data for Donor 2's CD8+ T cells (without use of Vivid dye), the variability in the estimates for $\mathrm{SD}\left[T_0^{div}\right]$ is substantially larger (IQR of 1.66 vs. 0.81) when fixing the parameter $\mathrm{E}\left[T_0^{div}\right]$. Returning to Figure 3.25, note that the scatter plots reveal strong correlation between $\mathrm{E}\left[T_0^{div}\right]$ and $\mathrm{SD}\left[T_0^{div}\right]$ in the case of Donor 1, but weak correlation (or no correlation) between these two parameters in the case of Donor 2. Therefore, the results when fixing one of the parameters in question are actually consistent with what one might expect.

Overall, comparing Figures 3.27 through 3.37 with Figures 3.10 through 3.21 (or, more precisely, comparing the IQRs for the corresponding box plots in those figures) reveals that fixing the value of $\mathrm{E}\left[T_0^{div}\right]$ is *not* a universally advantageous approach if our goal is to reduce variability in the parameter estimates. Interestingly, this approach is almost always advantageous in the case of Donor 1 data, but for many of the parameter estimates this approach causes an *increase* in variability when considering the Donor 2 data. For example, estimates for $\mathrm{E}\left[X_a\right]$, $\mathrm{SD}\left[X_a\right]$, $\mathrm{SD}\left[T_0^{div}\right]$, $\mathrm{E}\left[T^{die}\right]$, $\mathrm{SD}\left[T^{die}\right]$, and D_μ all experience significant increases in variability for at least some of the combinations of ViViD dye status and cell type when considering Donor 2 data.

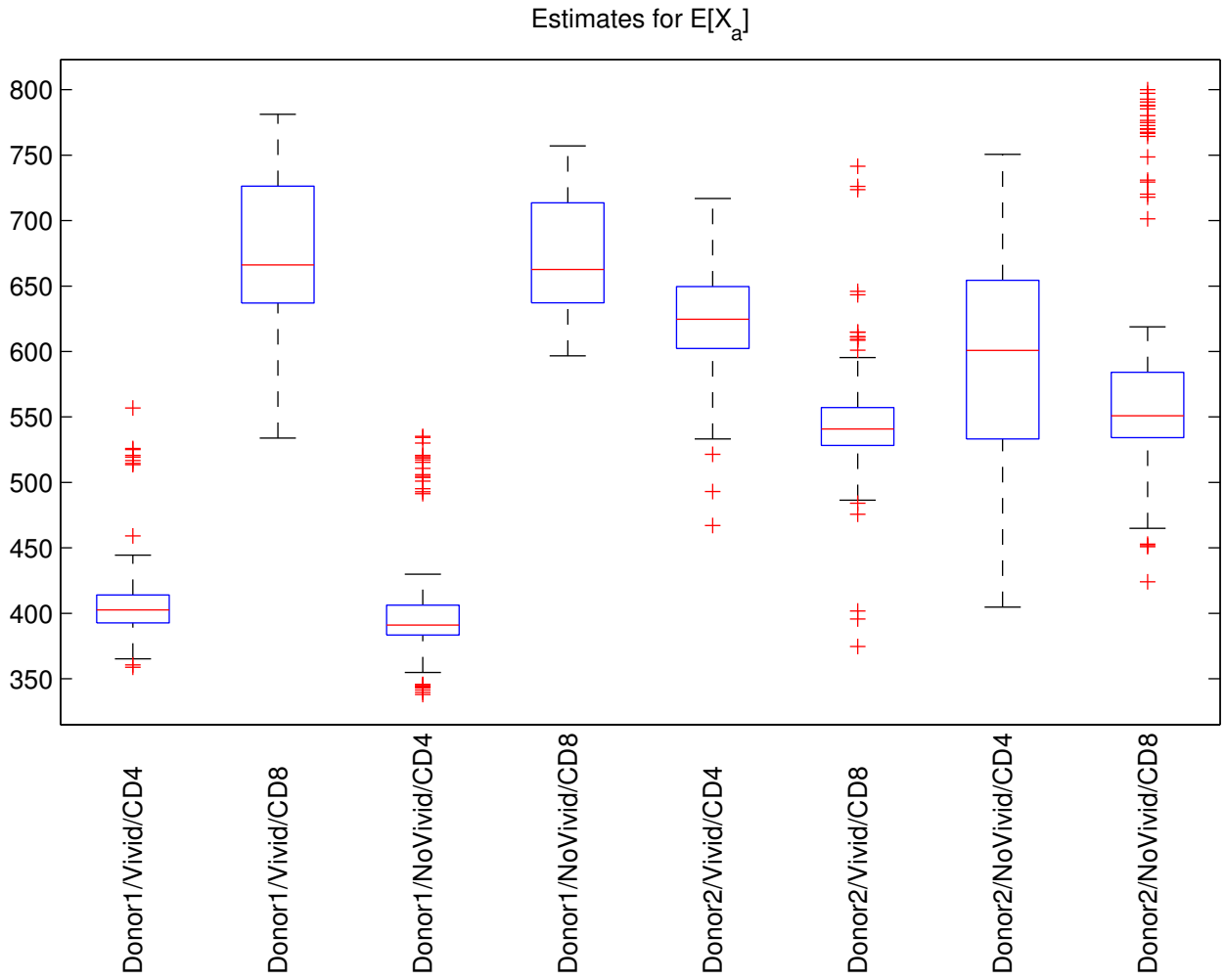


Figure 3.27: Box plots illustrating variability in estimates for the parameter $\mathrm{E}[X_a]$ when the parameter $\mathrm{E}\left[T_0^{div}\right]$ is fixed.

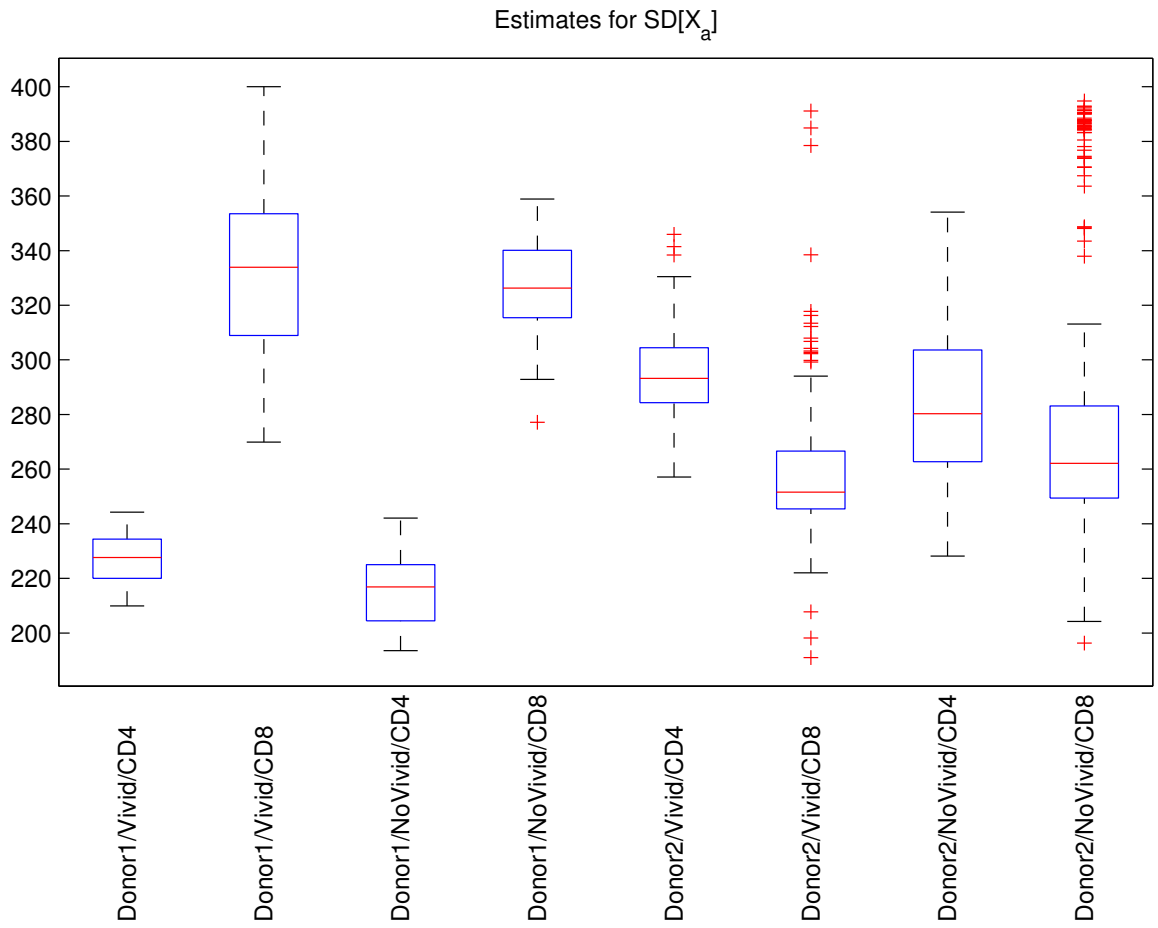


Figure 3.28: Box plots illustrating variability in estimates for the parameter SD $[X_a]$ when the parameter E $\left[T_0^{div}\right]$ is fixed.

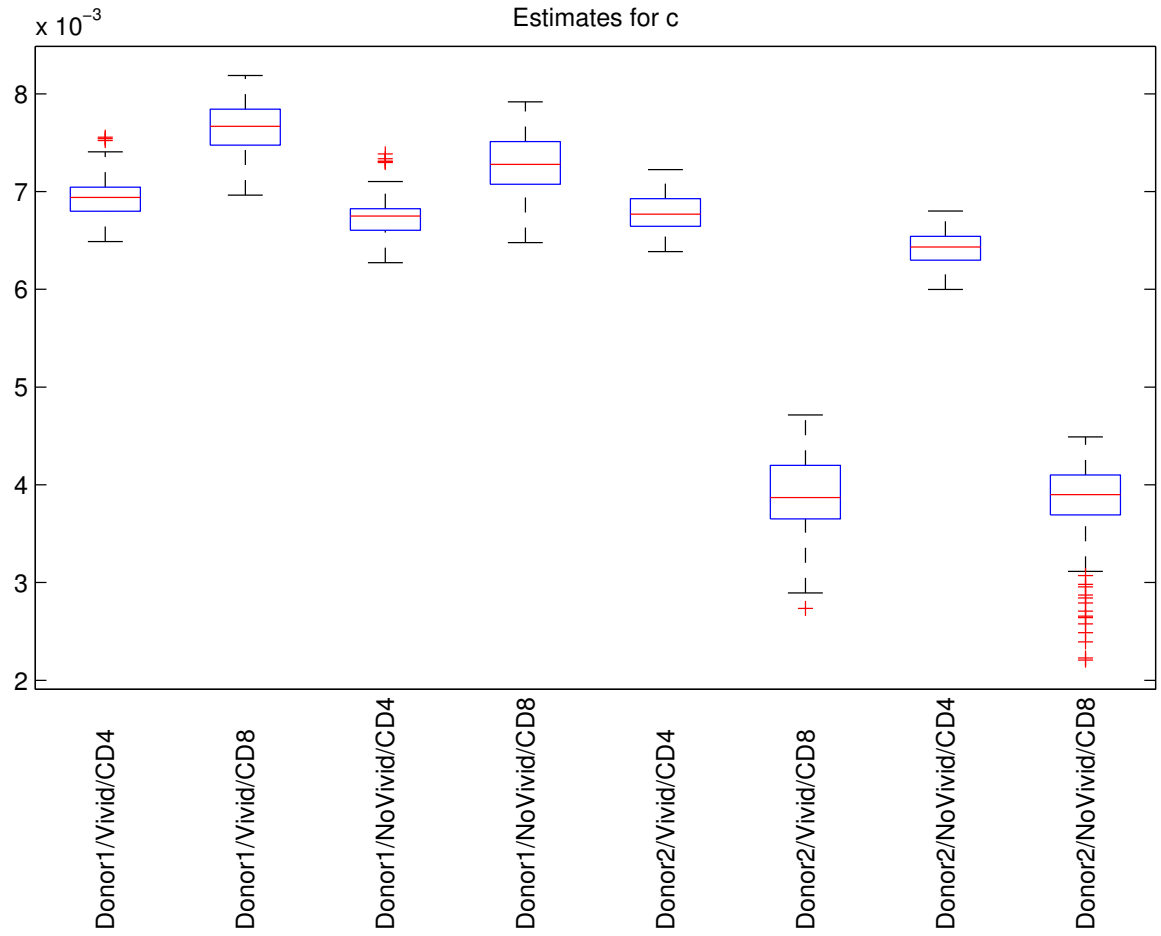


Figure 3.29: Box plots illustrating variability in estimates for the parameter c when the parameter $\mathrm{E}\left[T_0^{div}\right]$ is fixed.

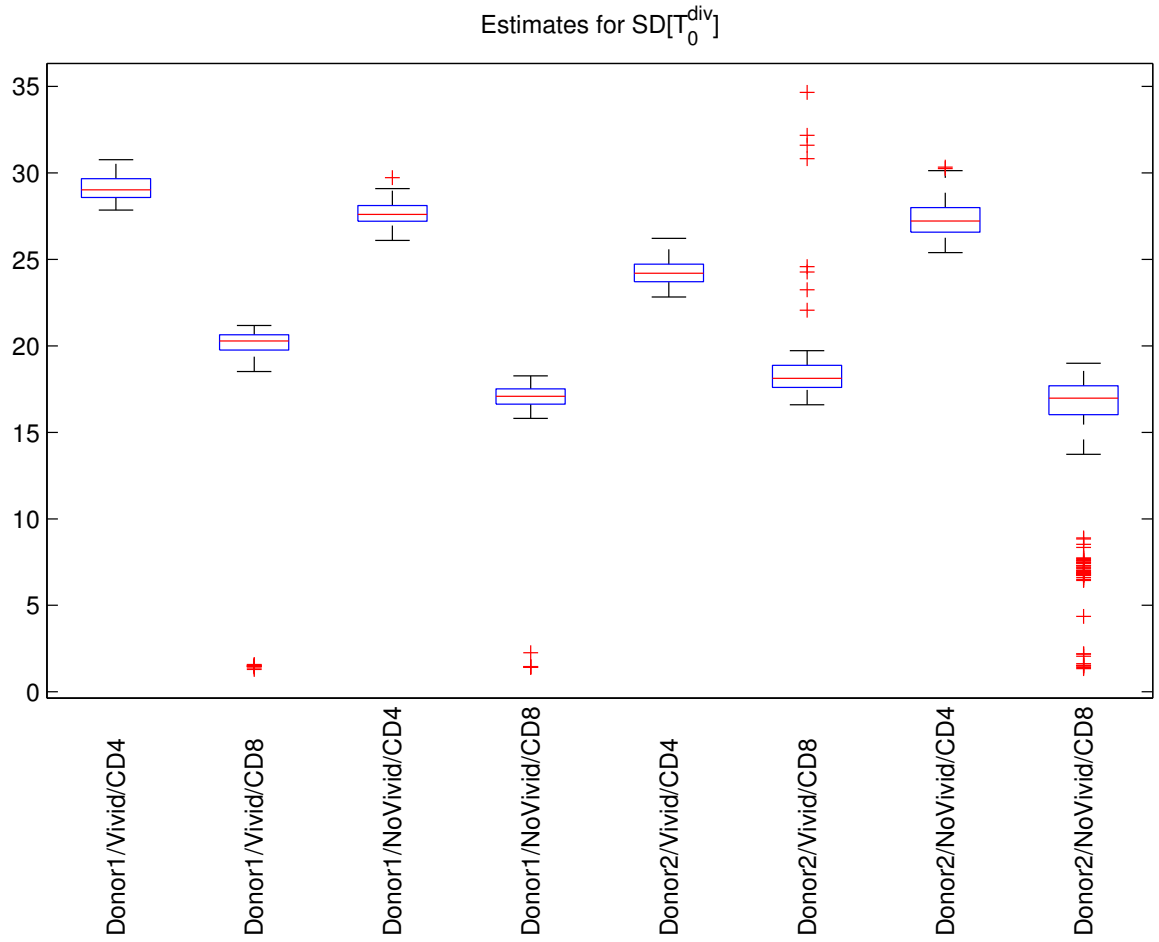


Figure 3.30: Box plots illustrating variability in estimates for the parameter SD $\left[T_0^{div}\right]$ when the parameter E $\left[T_0^{div}\right]$ is fixed.

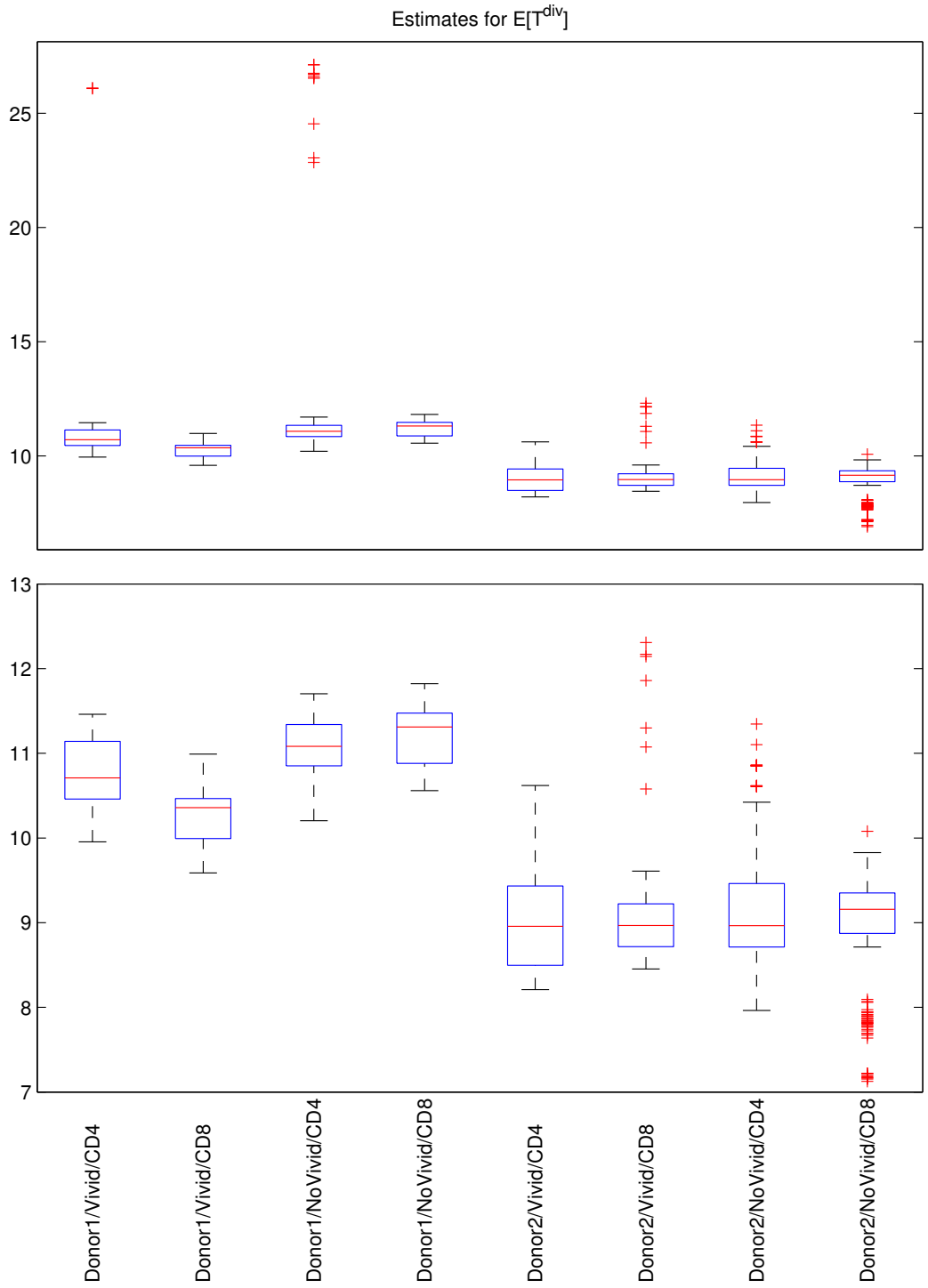


Figure 3.31: Box plots illustrating variability in estimates for the parameter E $\left[T^{div}\right]$ when the parameter E $\left[T_0^{div}\right]$ is fixed. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

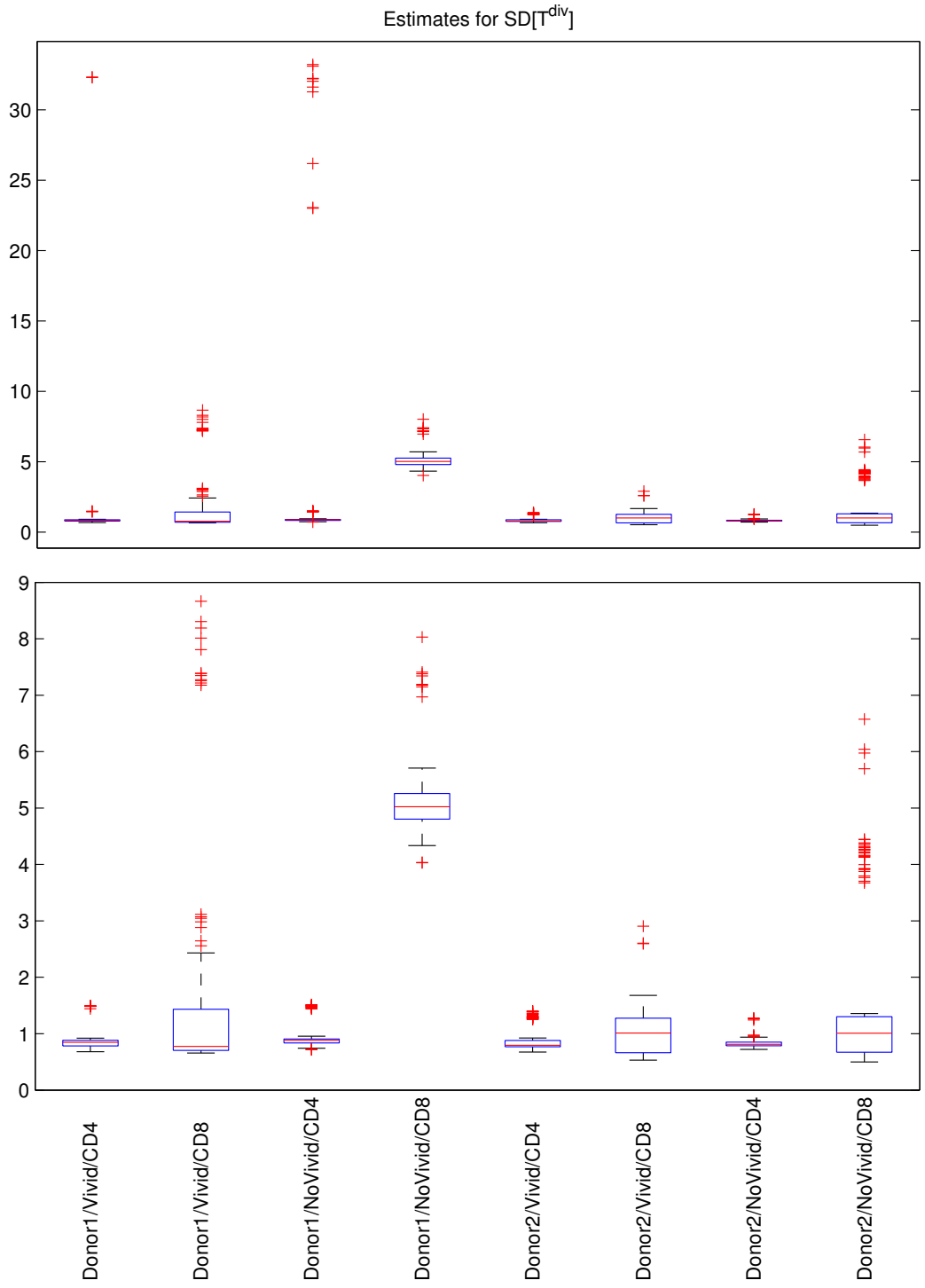


Figure 3.32: Box plots illustrating variability in estimates for the parameter SD $\left[T^{div}\right]$ when the parameter E $\left[T_0^{div}\right]$ is fixed. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

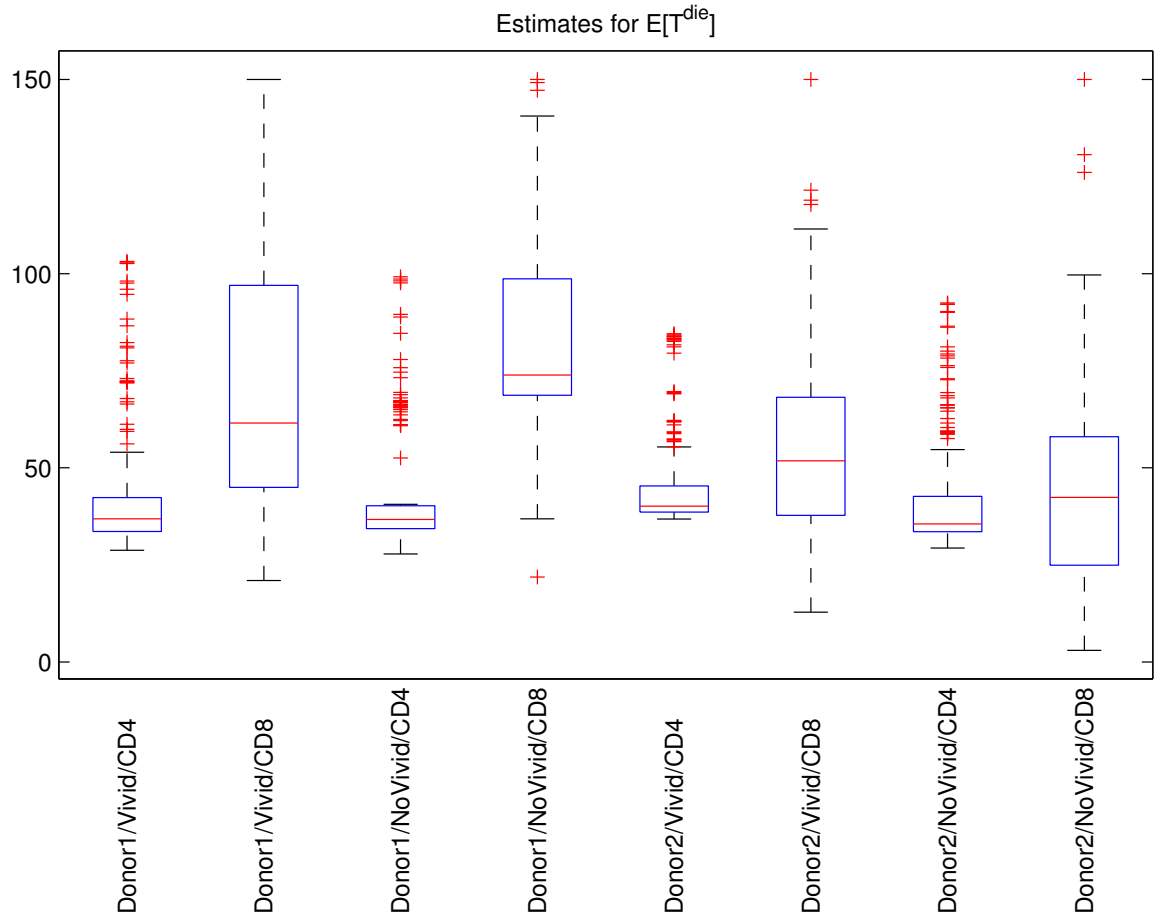


Figure 3.33: Box plots illustrating variability in estimates for the parameter E $\left[T^{die}\right]$ when the parameter E $\left[T_0^{div}\right]$ is fixed.

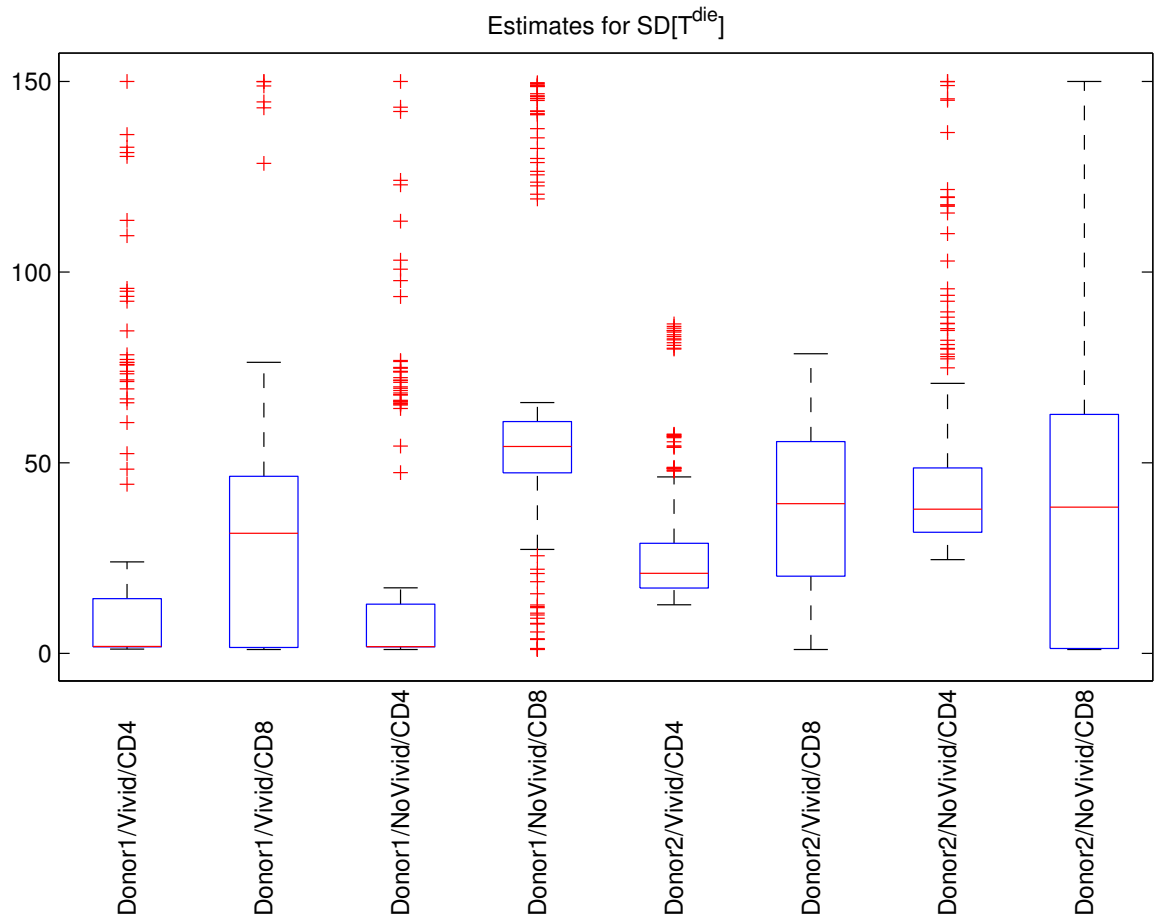


Figure 3.34: Box plots illustrating variability in estimates for the parameter SD $\left[T^{die}\right]$ when the parameter E $\left[T_0^{div}\right]$ is fixed.

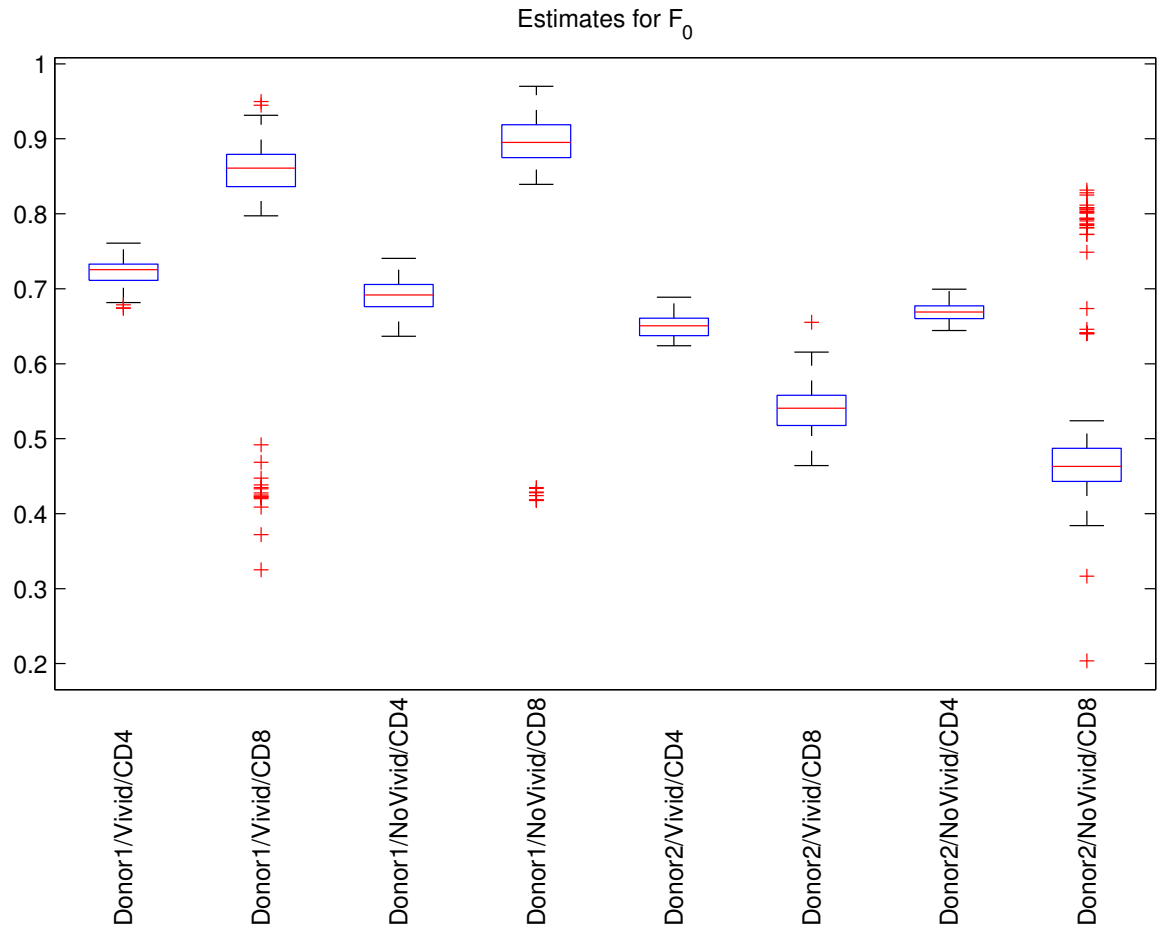


Figure 3.35: Box plots illustrating variability in estimates for the parameter F_0 when the parameter $\mathrm{E}\left[T_0^{div}\right]$ is fixed.

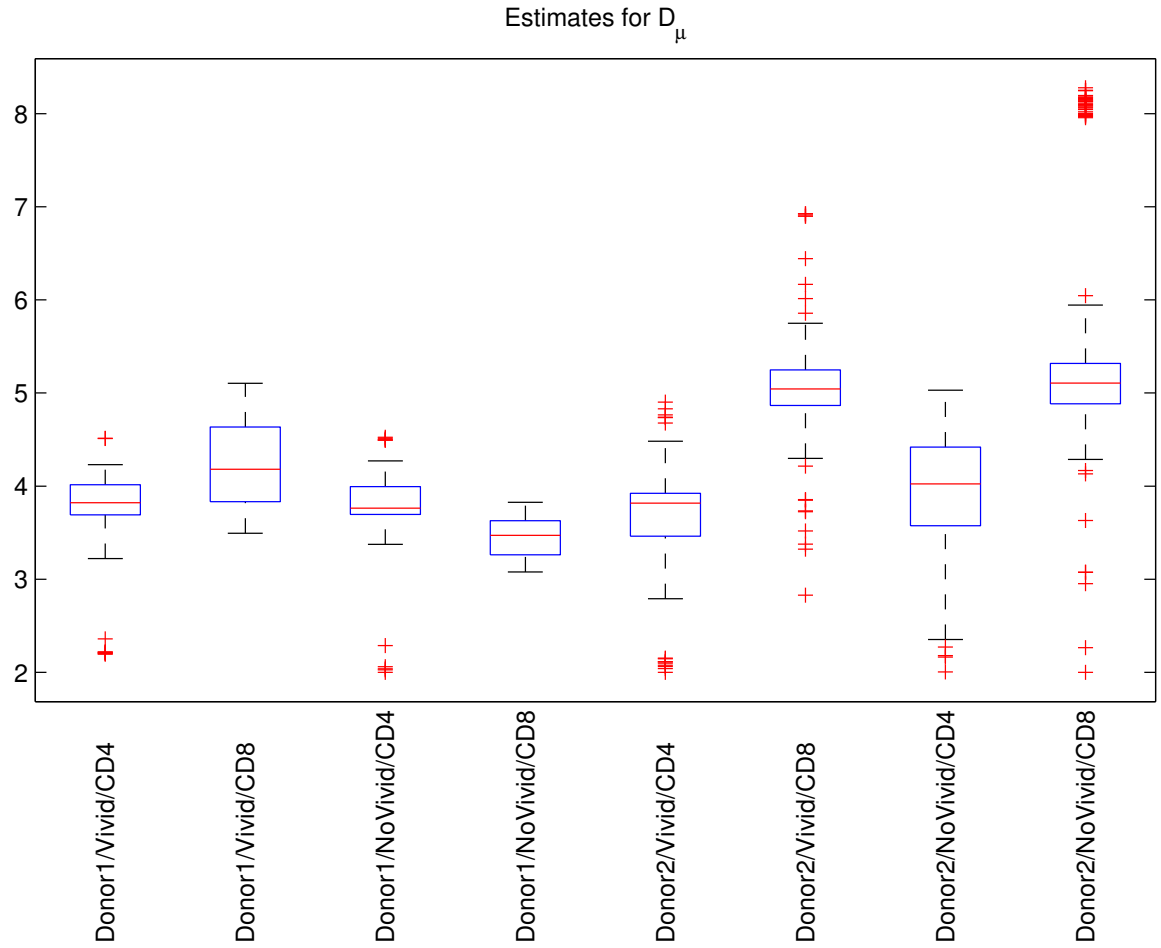


Figure 3.36: Box plots illustrating variability in estimates for the parameter D_μ when the parameter $\mathrm{E}\left[T_0^{div}\right]$ is fixed.

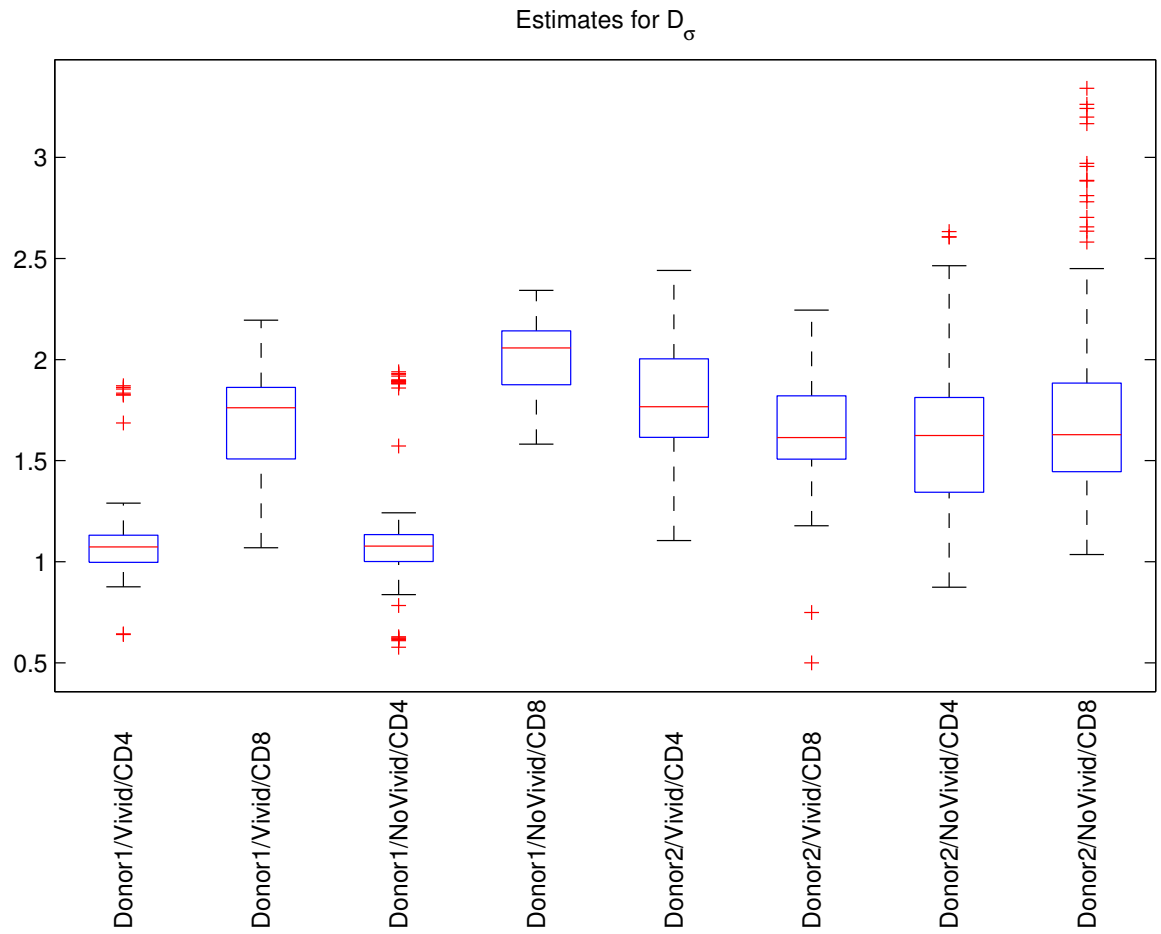


Figure 3.37: Box plots illustrating variability in estimates for the parameter D_σ when the parameter E $\left[T_0^{div}\right]$ is fixed.

In Figures 3.38 through 3.48, we show the results of applying our modified parameter estimation technique in case (ii). Again, note that there is one figure for each of the 11 model parameters that is not fixed. Recall that our goal in fixing the value of the parameter $\mathrm{E}\left[T^{die}\right]$ is to reduce the variability in the parameter estimates for $\mathrm{SD}\left[T^{die}\right]$, which seems to be correlated with $\mathrm{E}\left[T^{die}\right]$ (based on Figure 3.26). Comparing Figures 3.18 and 3.45, we see that using a fixed value for $\mathrm{E}\left[T^{die}\right]$ does considerably reduce the variability in the estimates of $\mathrm{SD}\left[T^{die}\right]$ in most cases; however, when using data for Donor 2's CD4+ T cells (with use of Vivid dye), the variability in the estimates for $\mathrm{SD}\left[T^{die}\right]$ is substantially larger (IQR of 10.84 vs. 5.54) when fixing the parameter $\mathrm{E}\left[T^{die}\right]$. It is not clear from Figure 3.26 why there should be such a discrepancy in this particular case.

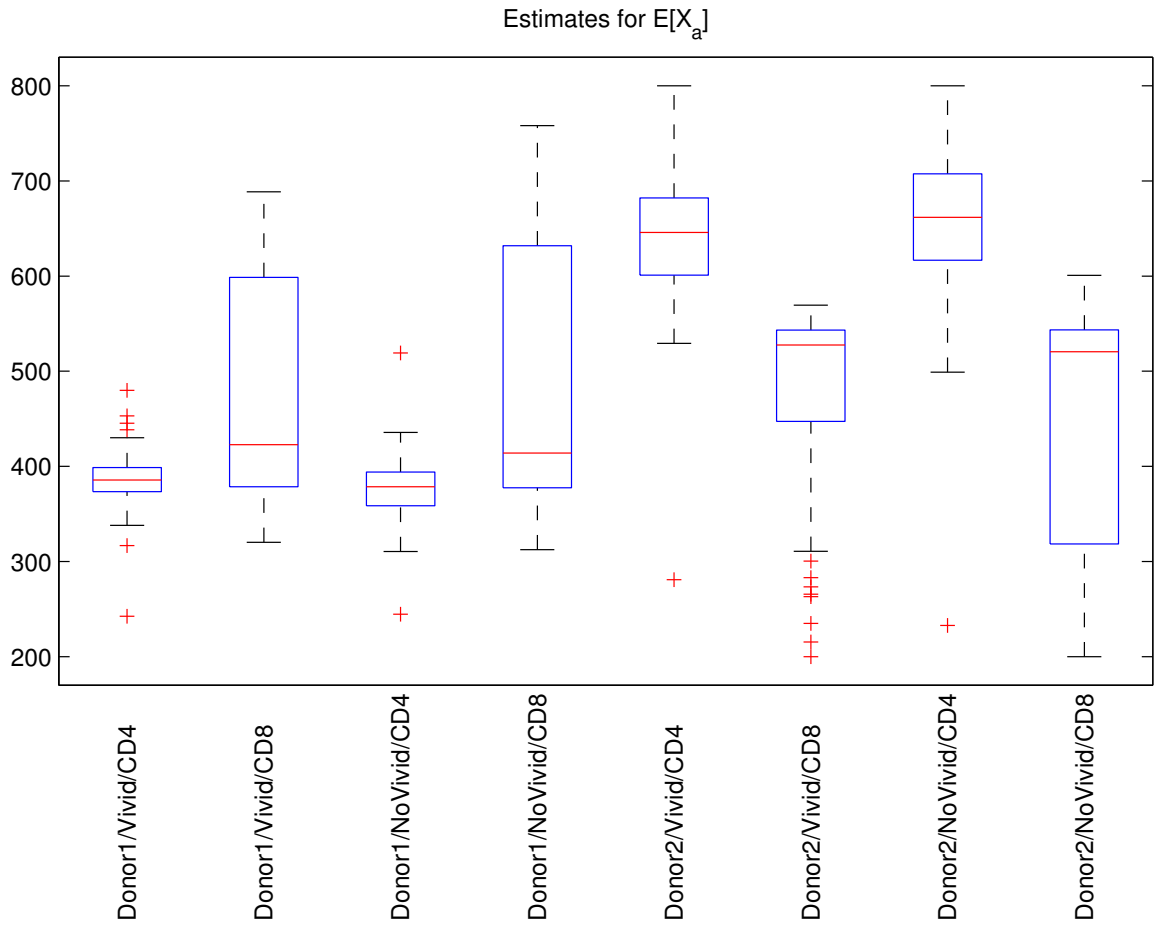


Figure 3.38: Box plots illustrating variability in estimates for the parameter $\mathrm{E}\left[X_a\right]$ when the parameter $\mathrm{E}\left[T^{die}\right]$ is fixed.

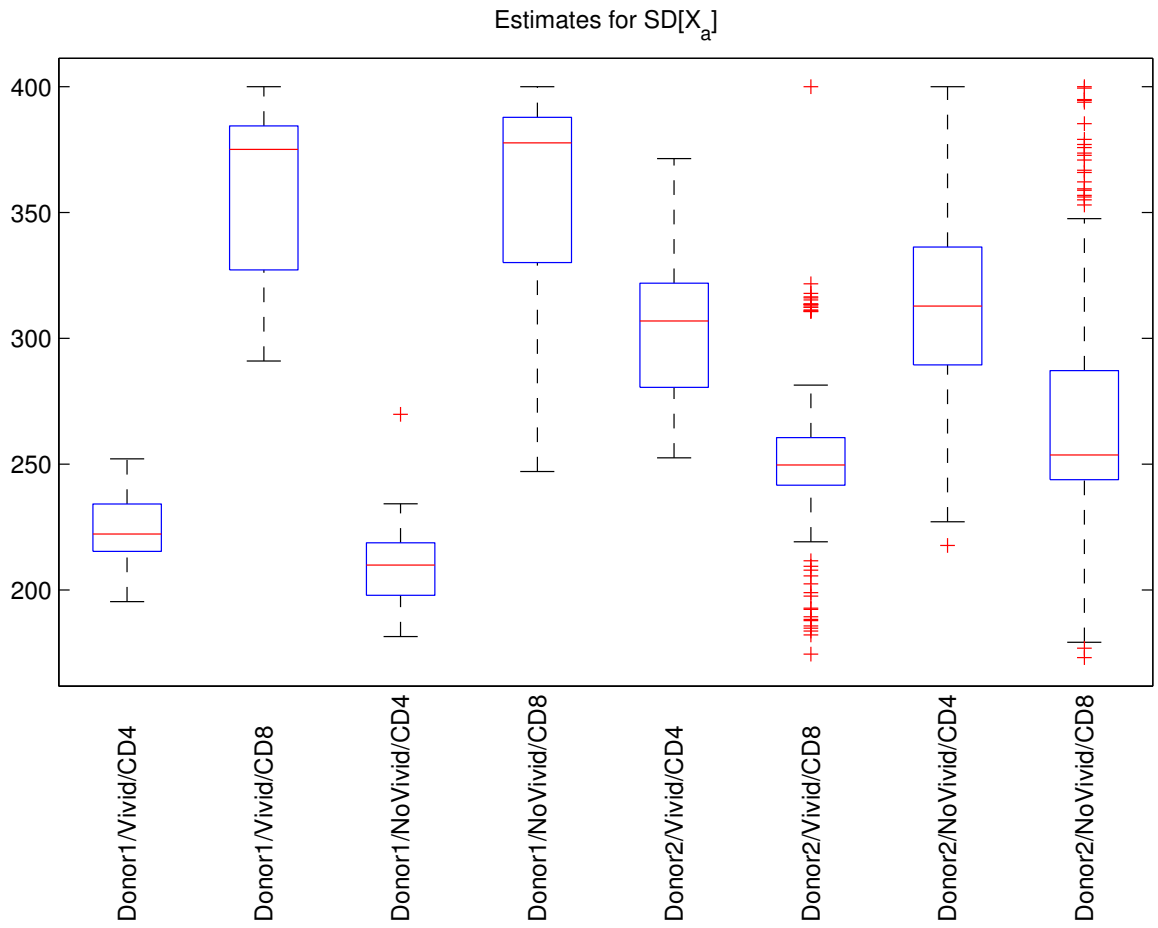


Figure 3.39: Box plots illustrating variability in estimates for the parameter SD $[X_a]$ when the parameter E $\left[T^{die}\right]$ is fixed.

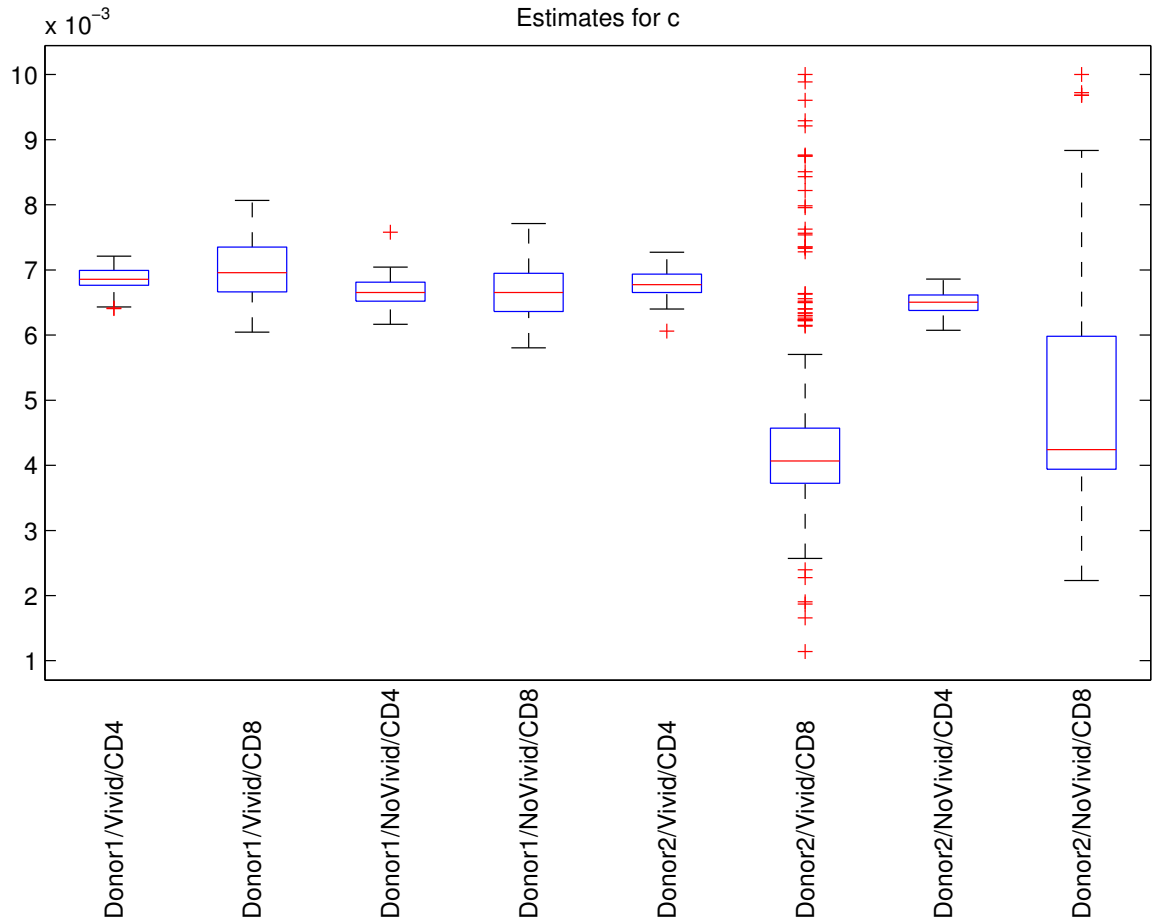


Figure 3.40: Box plots illustrating variability in estimates for the parameter c when the parameter $\mathrm{E}\left[T^{die}\right]$ is fixed.

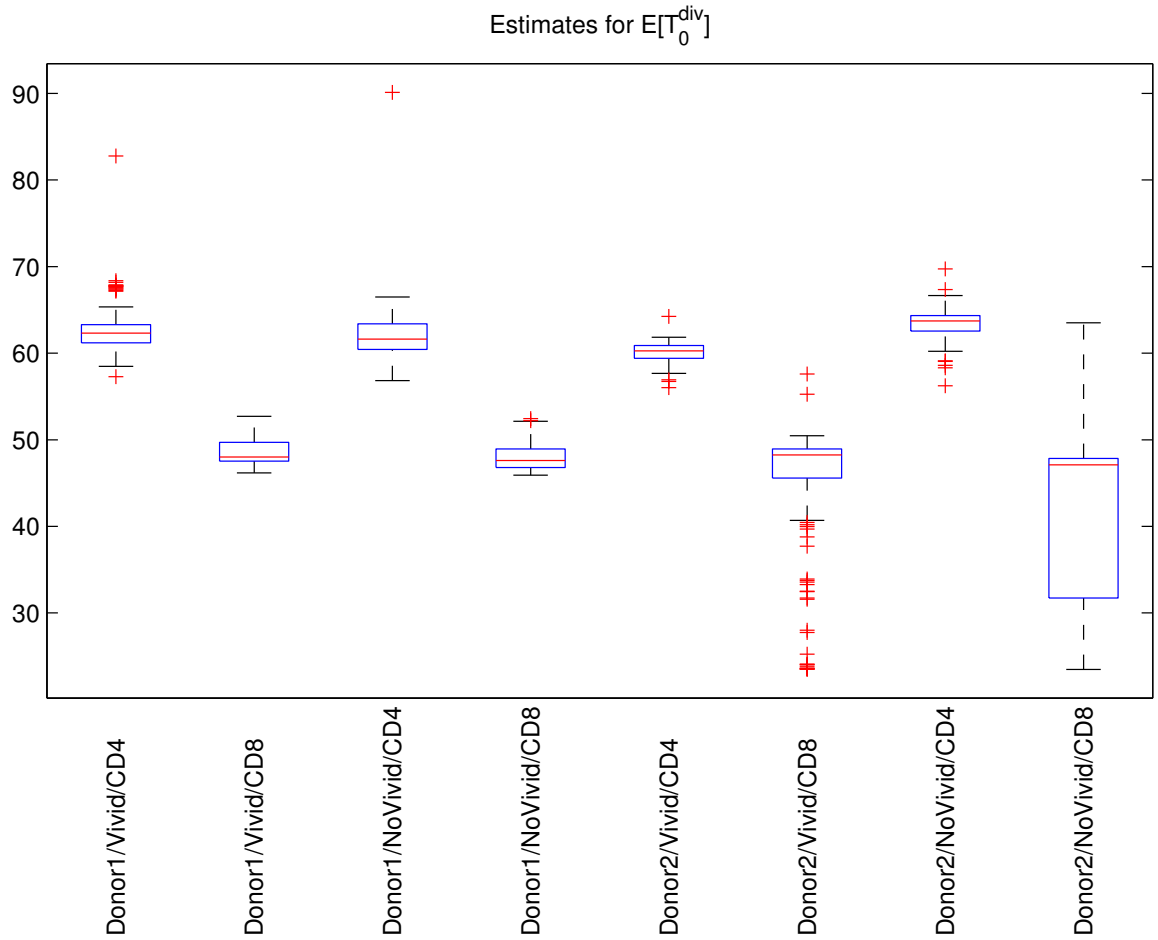


Figure 3.41: Box plots illustrating variability in estimates for the parameter E $\left[T_0^{div}\right]$ when the parameter E $\left[T^{die}\right]$ is fixed.

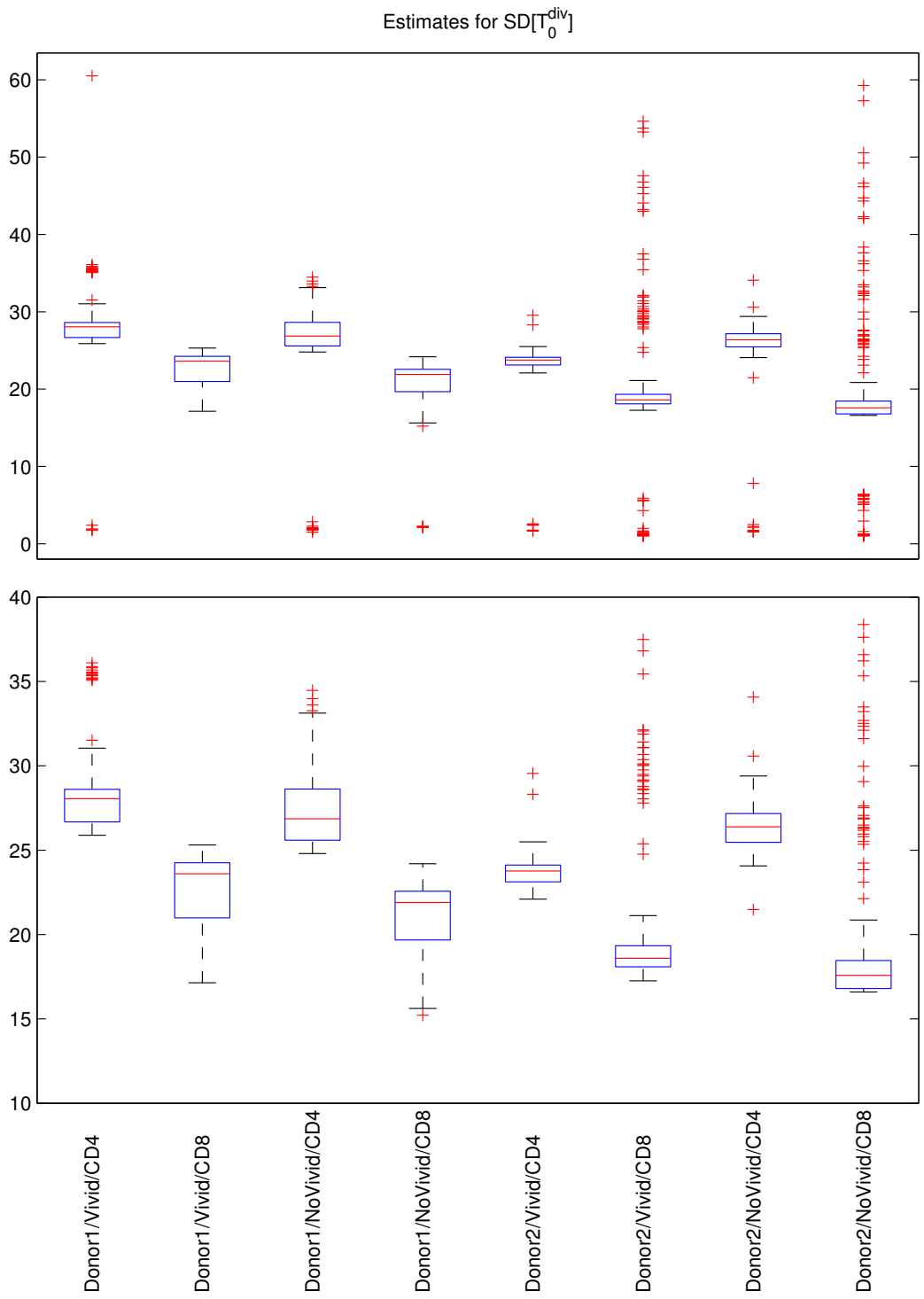


Figure 3.42: Box plots illustrating variability in estimates for the parameter SD $\left[T_0^{div}\right]$ when the parameter E $\left[T^{die}\right]$ is fixed. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

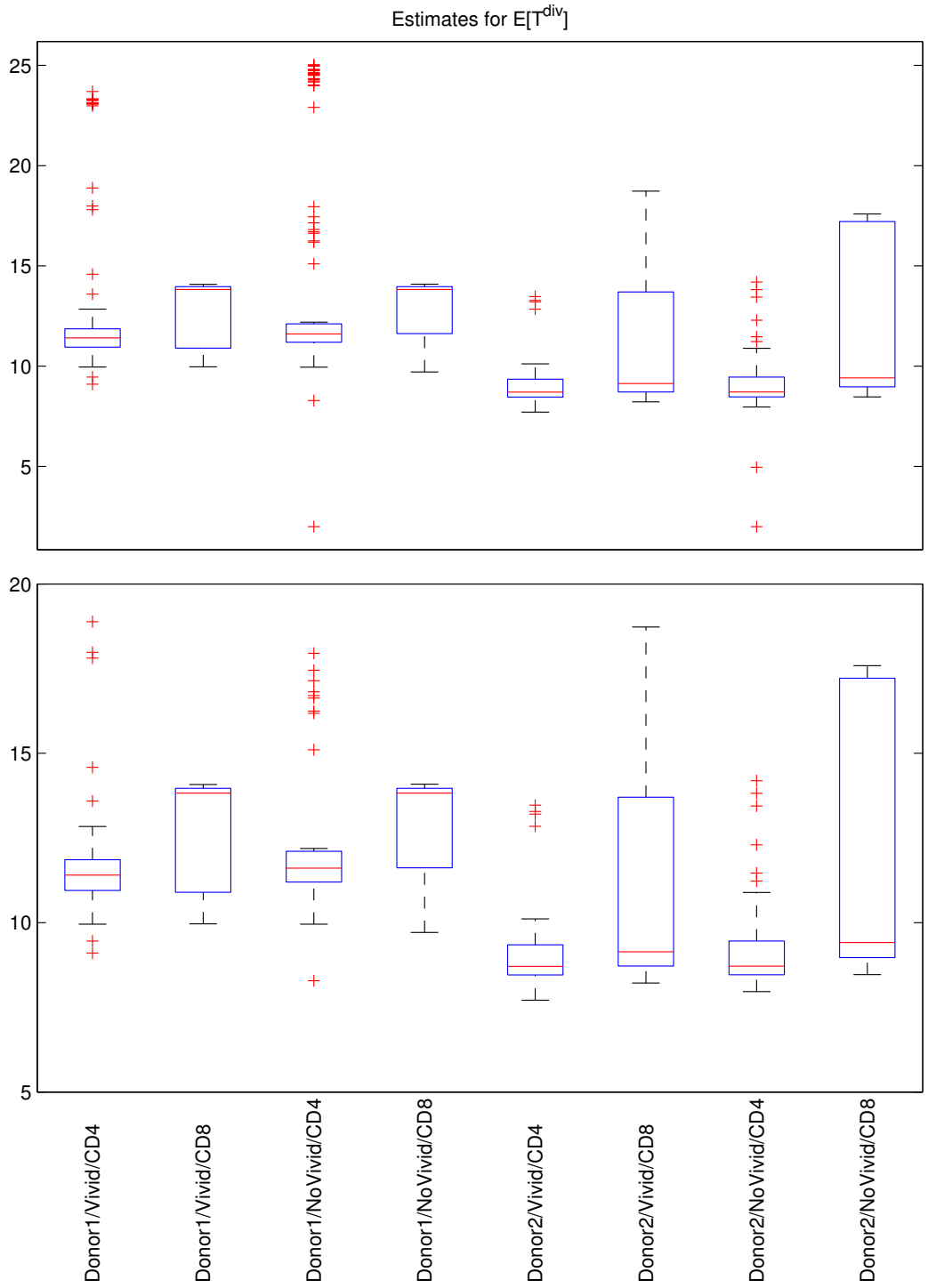


Figure 3.43: Box plots illustrating variability in estimates for the parameter E $\left[T^{div}\right]$ when the parameter E $\left[T^{die}\right]$ is fixed. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

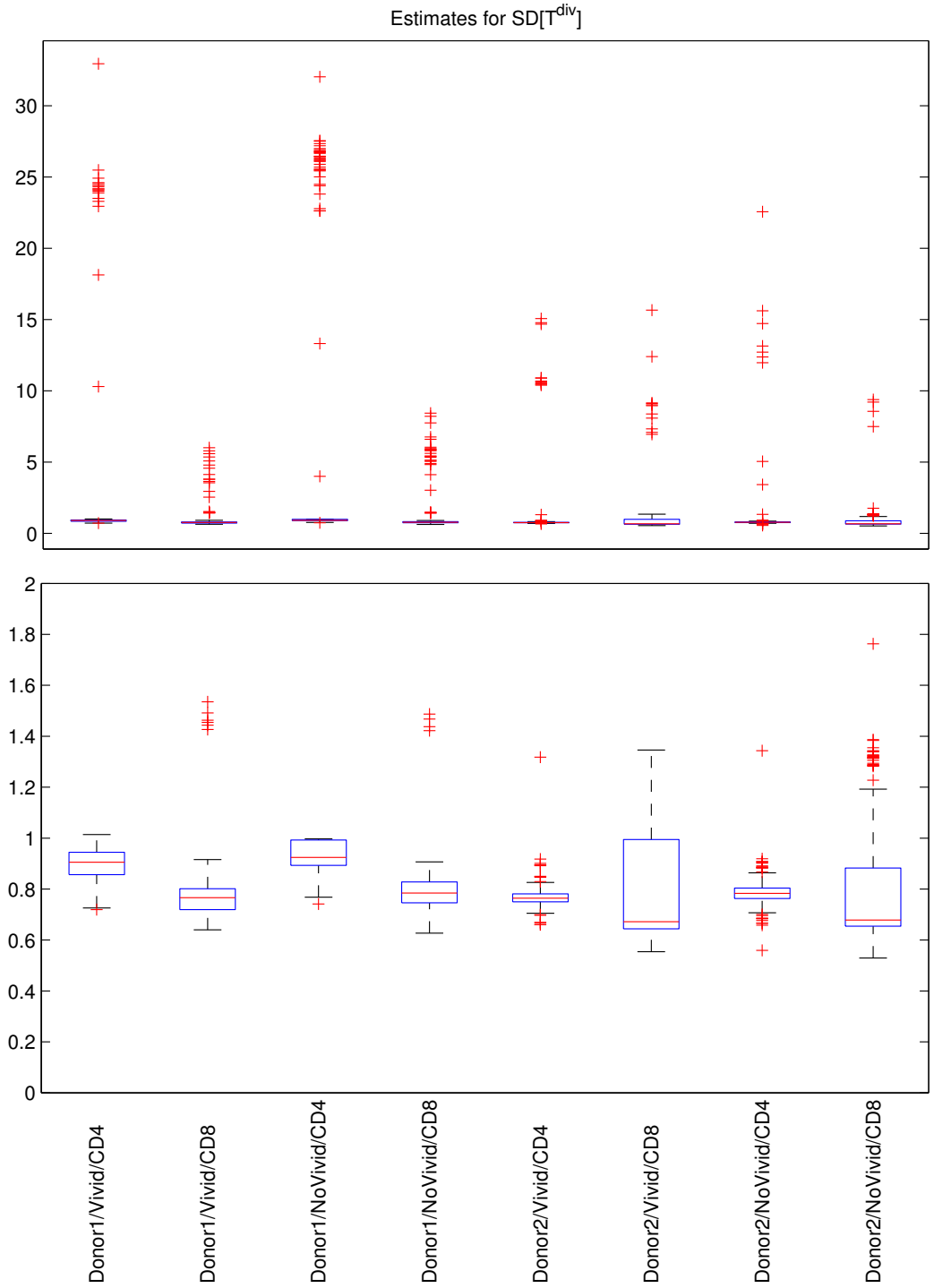


Figure 3.44: Box plots illustrating variability in estimates for the parameter SD $\left[T^{div}\right]$ when the parameter E $\left[T^{die}\right]$ is fixed. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

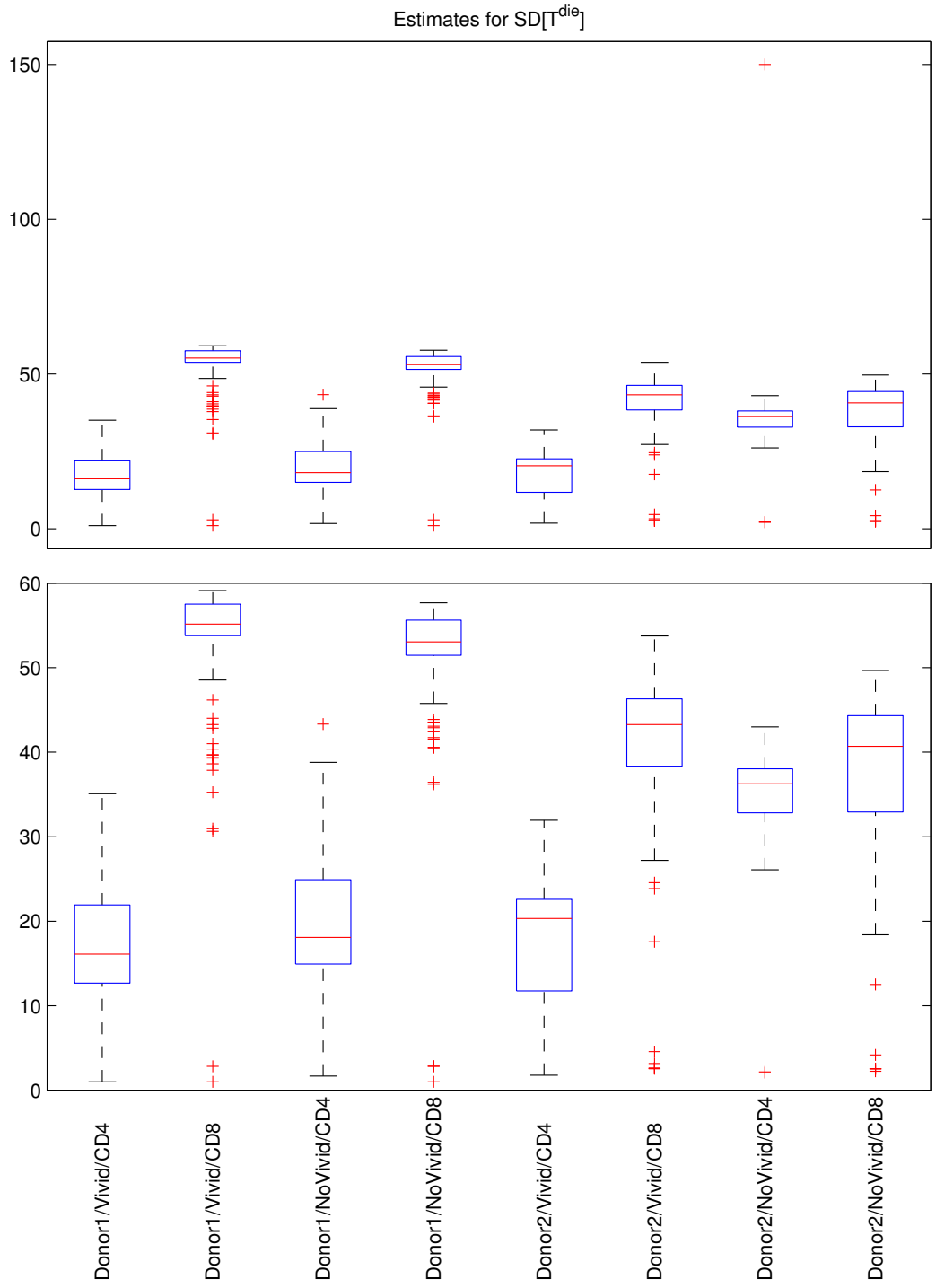


Figure 3.45: Box plots illustrating variability in estimates for the parameter SD $\left[T^{die}\right]$ when the parameter E $\left[T^{die}\right]$ is fixed. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

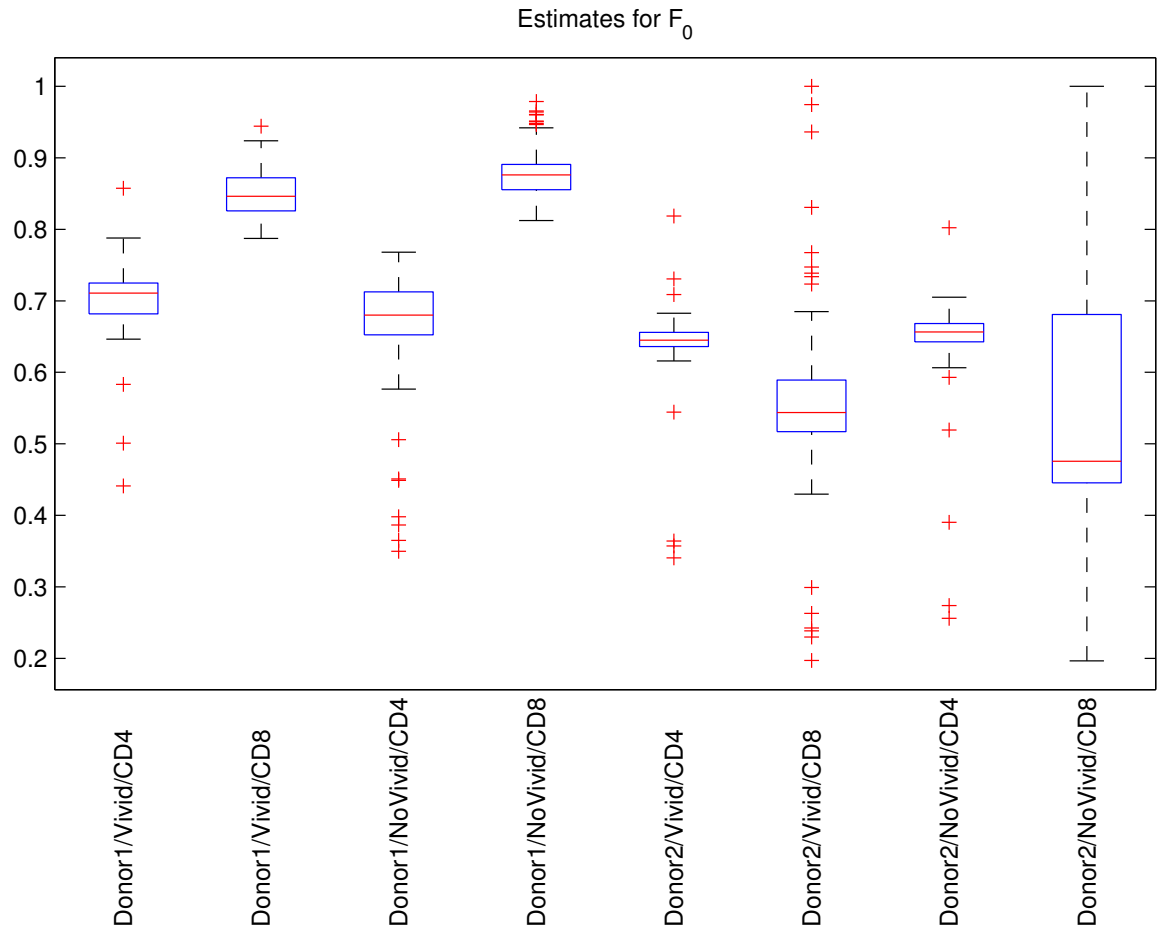


Figure 3.46: Box plots illustrating variability in estimates for the parameter F_0 when the parameter $\mathrm{E}\left[T^{die}\right]$ is fixed.

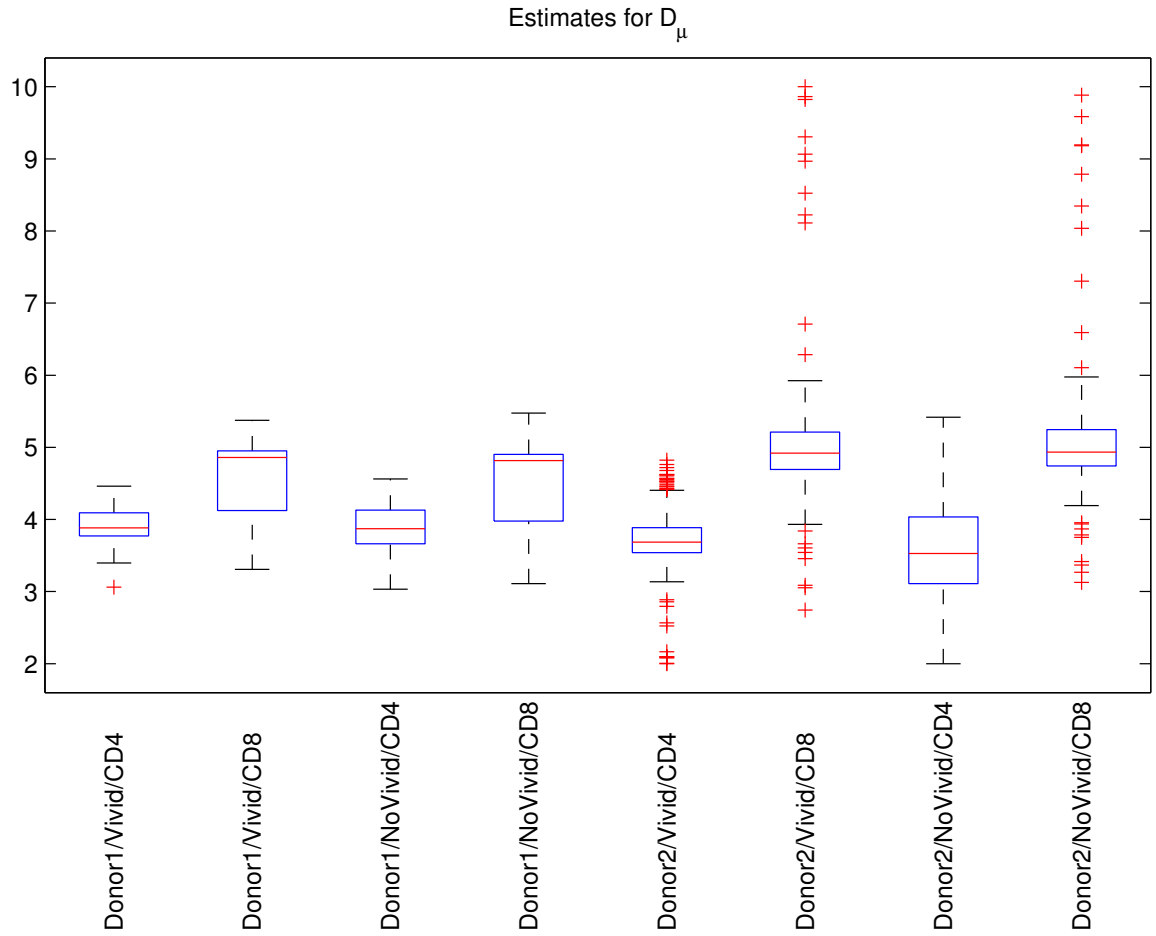


Figure 3.47: Box plots illustrating variability in estimates for the parameter D_μ when the parameter $\mathrm{E}\left[T^{die}\right]$ is fixed.

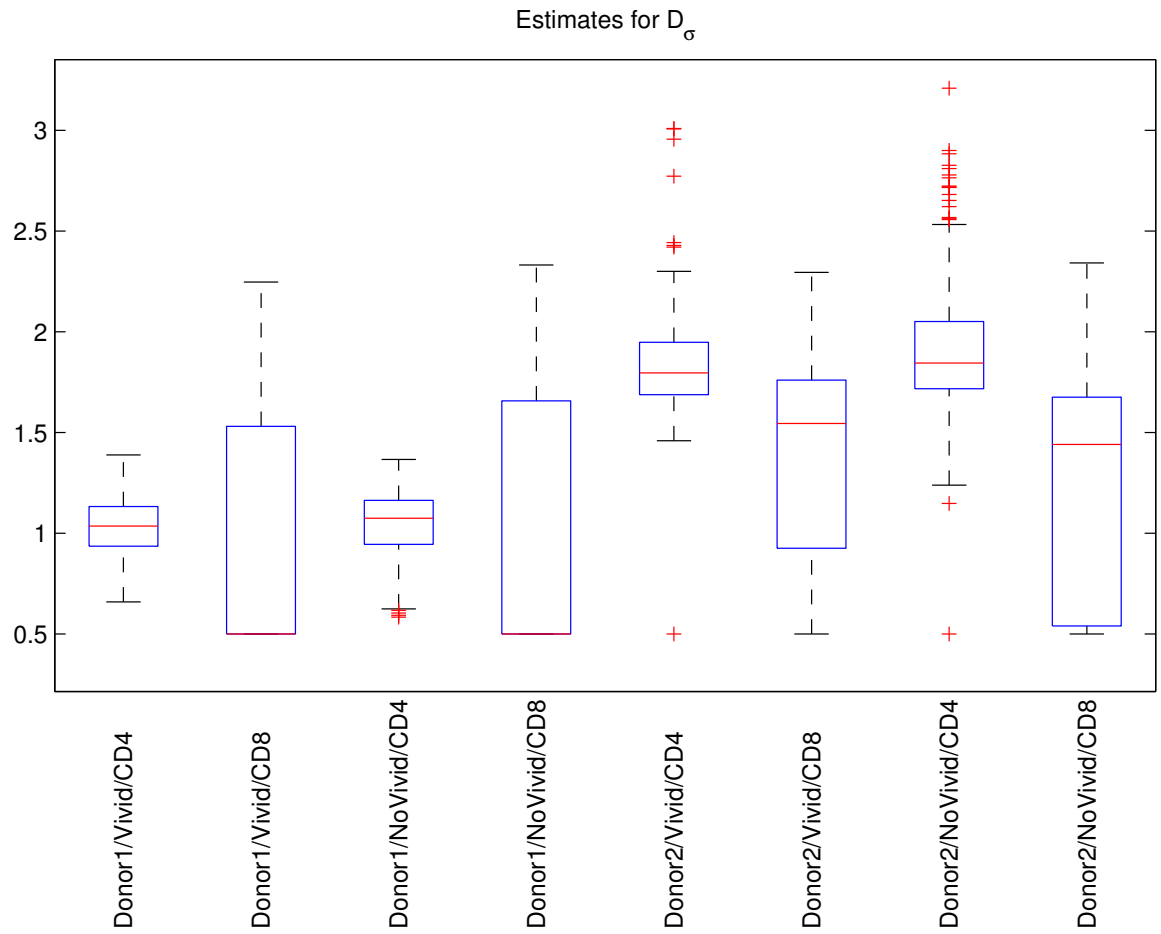


Figure 3.48: Box plots illustrating variability in estimates for the parameter D_σ when the parameter $\mathrm{E}\left[T^{die}\right]$ is fixed.

If we compare Figures 3.38 through 3.48 with Figures 3.10 through 3.21, it is evident that fixing the value of $\mathrm{E}\left[T^{die}\right]$ causes increases in the variability of parameter estimates in many cases. Unlike the situation with the previous fixed parameter ($\mathrm{E}\left[T_0^{div}\right]$), these increases are prevalent when considering Donor 1 data *and* Donor 2 data. We are therefore forced to conclude that fixing the value of $\mathrm{E}\left[T^{die}\right]$ also fails as a universally advantageous scheme for reducing variability in the parameter estimates.

3.3.4 Parameter Estimates Obtained Using Only Data for Days 1 through 3

As was demonstrated in Section 3.2, the amount of relative variation in the cell counts undergoes significant changes between Day 3 and Day 5. We therefore also attempted to estimate parameters using only data from Days 1 through 3. The results of applying our parameter estimation technique to such reduced data sets are provided in Figures 3.49 through 3.60. Since in this case we have triplicate samples for each of *three* days, there are $3^3 = 27$ possible ways to form a three-day time series data set for each donor. Therefore, each box plot in the above-referenced figures represents a summary of 27 parameter estimates.

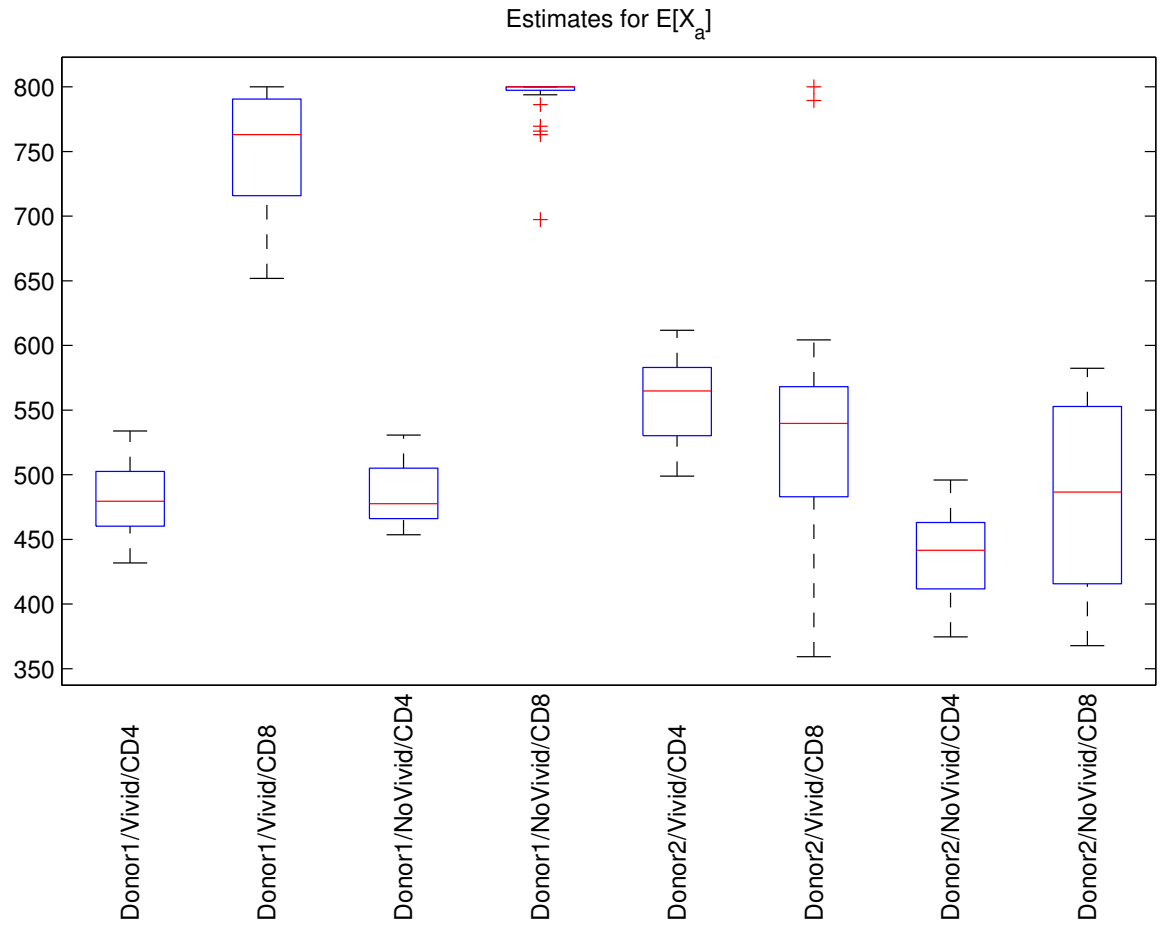


Figure 3.49: Box plots illustrating variability in estimates for the parameter $\mathrm{E}[X_a]$ when using only data from Days 1 through 3.

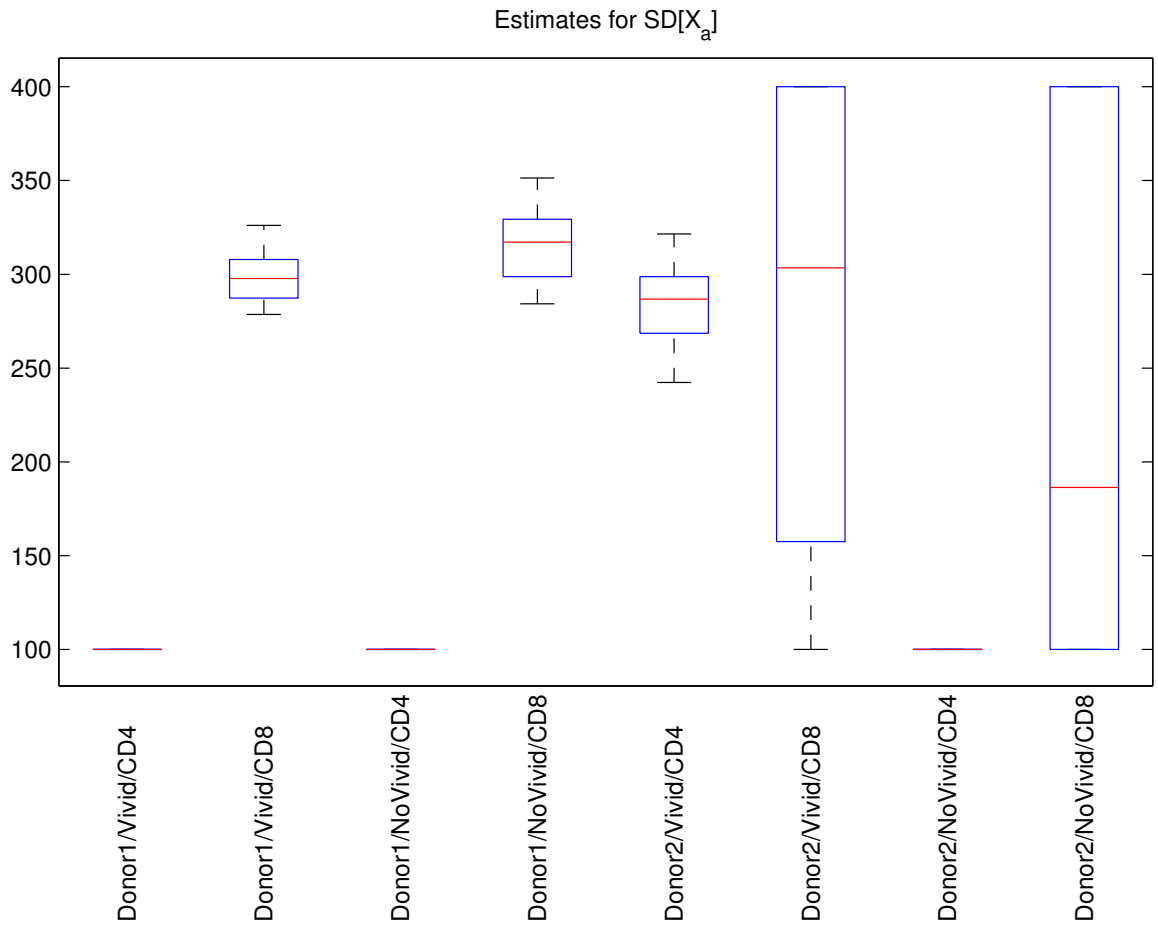


Figure 3.50: Box plots illustrating variability in estimates for the parameter SD $[X_a]$ when using only data from Days 1 through 3.

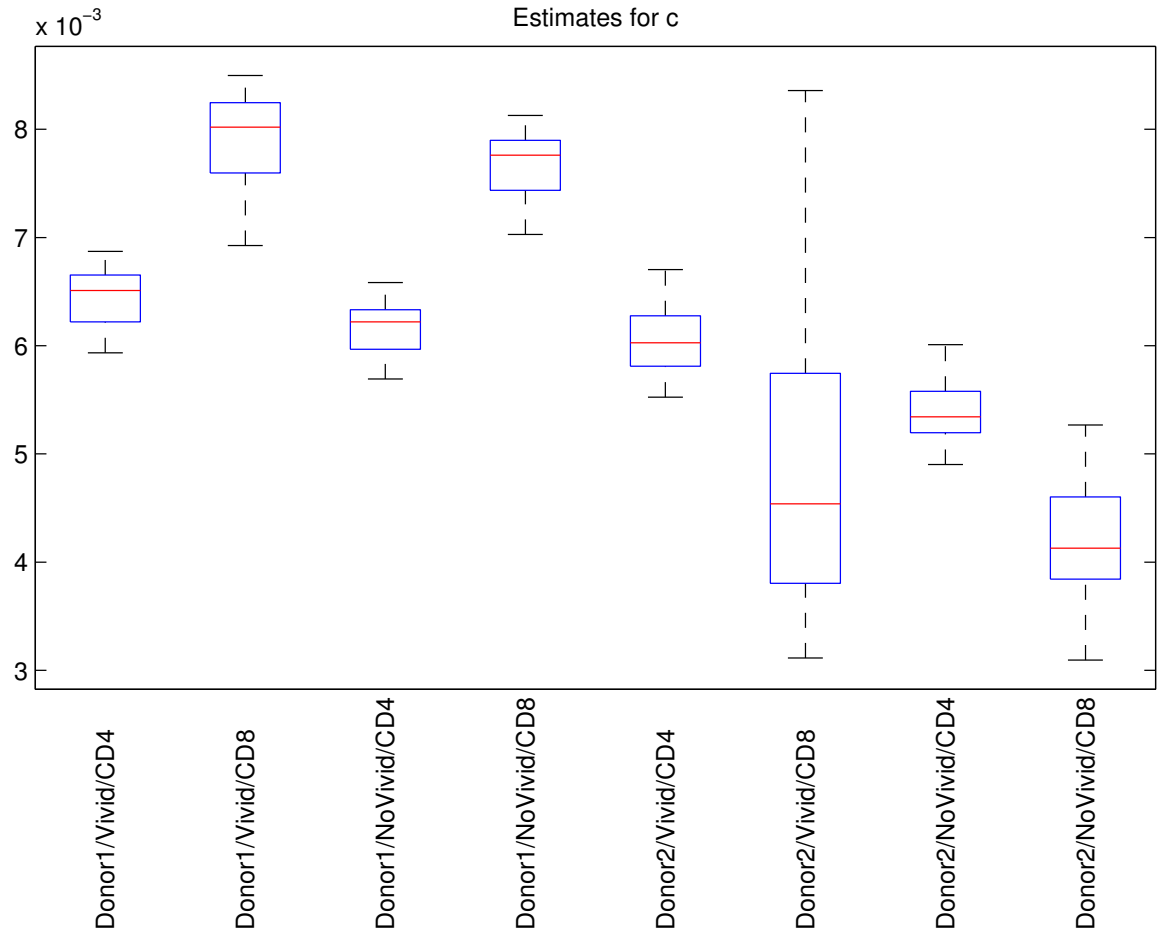


Figure 3.51: Box plots illustrating variability in estimates for the parameter c when using only data from Days 1 through 3.

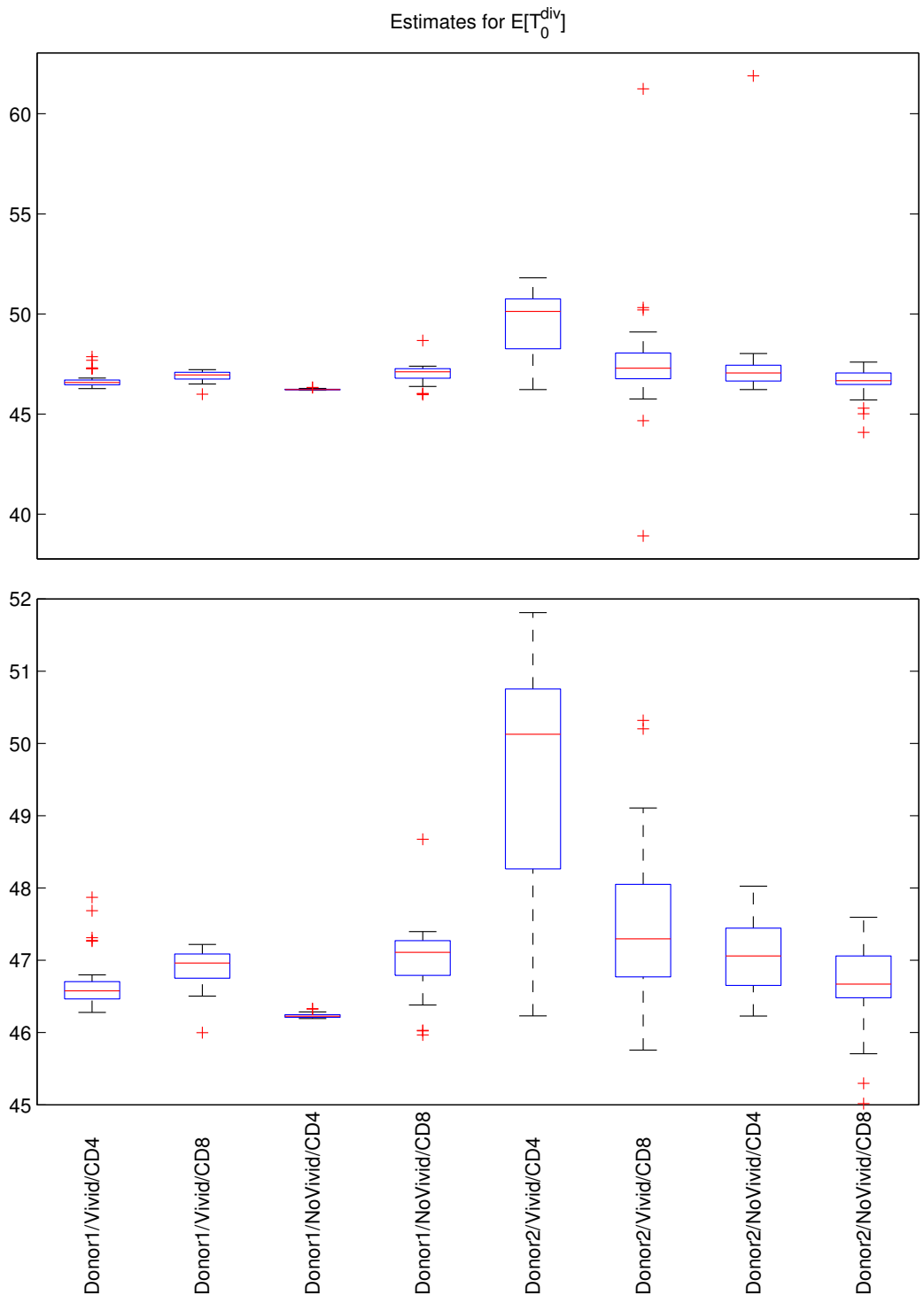


Figure 3.52: Box plots illustrating variability in estimates for the parameter $\mathrm{E}\left[T_0^{div}\right]$ when using only data from Days 1 through 3. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

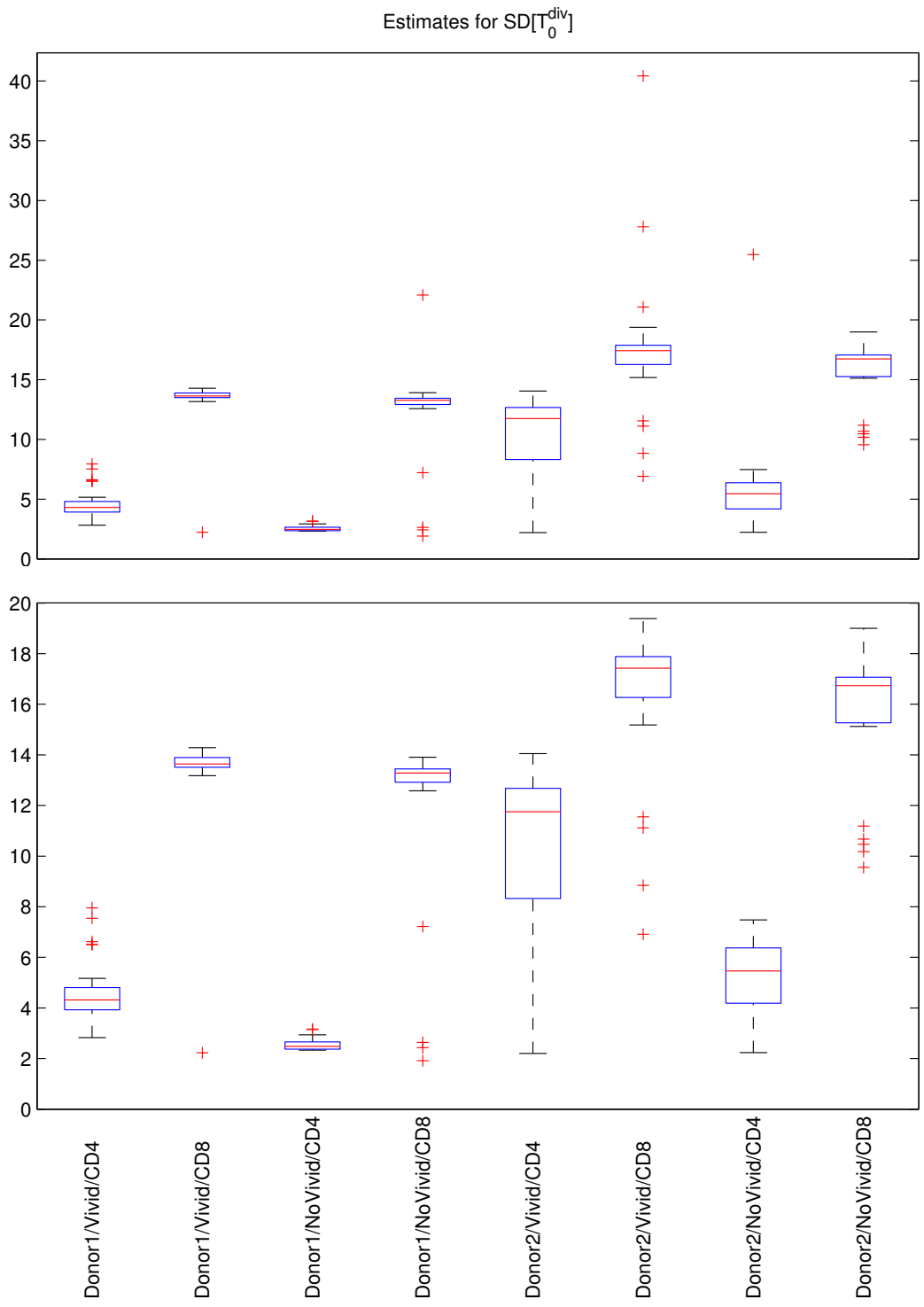


Figure 3.53: Box plots illustrating variability in estimates for the parameter SD $\left[T_0^{div}\right]$ when using only data from Days 1 through 3. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

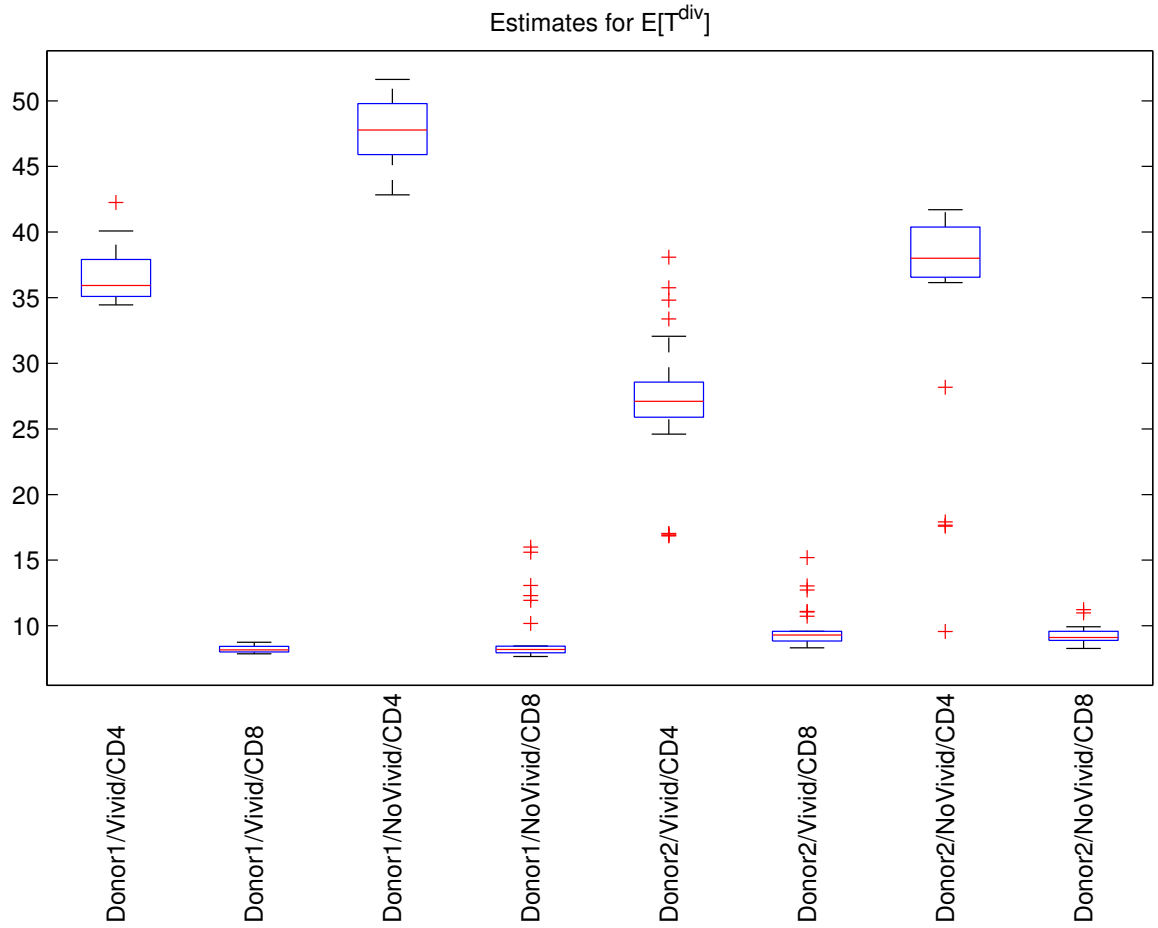


Figure 3.54: Box plots illustrating variability in estimates for the parameter E $\left[T^{div}\right]$ when using only data from Days 1 through 3.

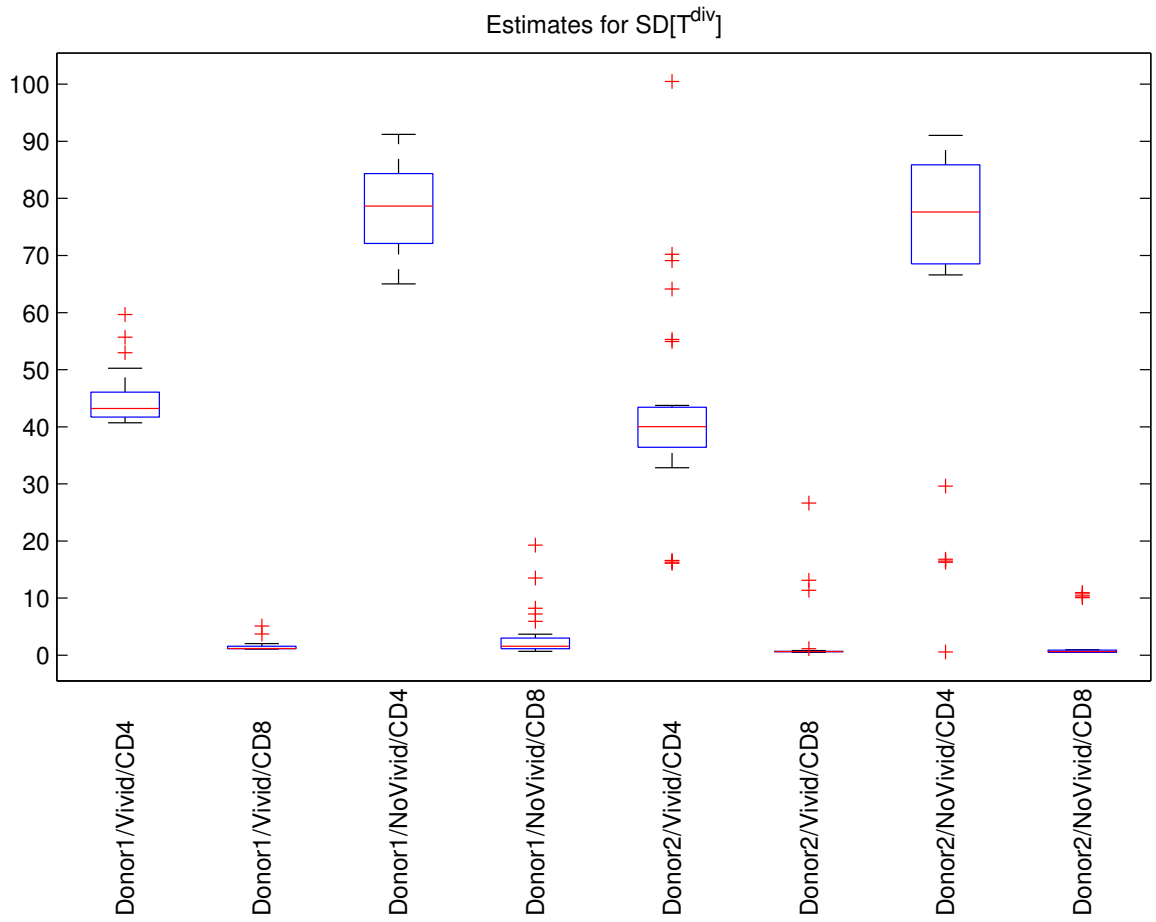


Figure 3.55: Box plots illustrating variability in estimates for the parameter SD $\left[T^{div}\right]$ when using only data from Days 1 through 3.

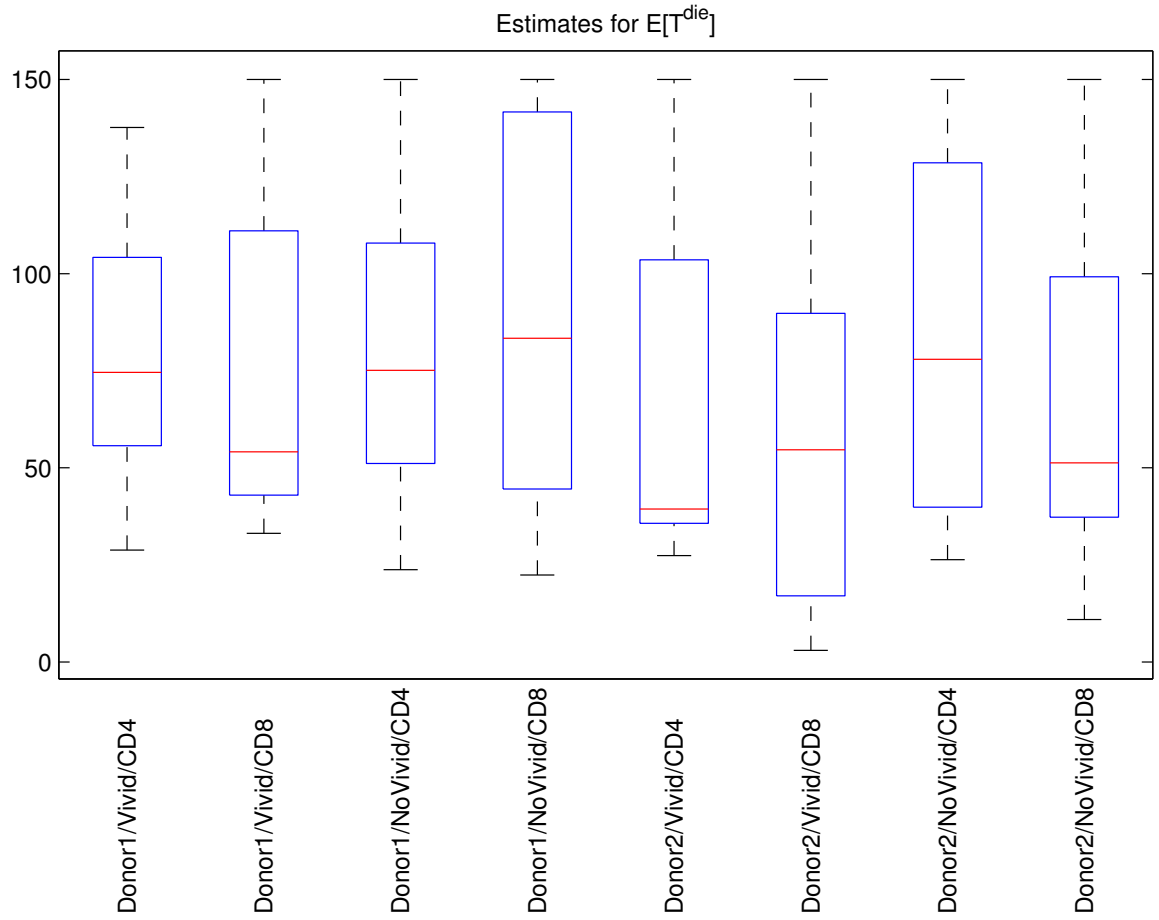


Figure 3.56: Box plots illustrating variability in estimates for the parameter $\mathrm{E}\left[T^{die}\right]$ when using only data from Days 1 through 3.

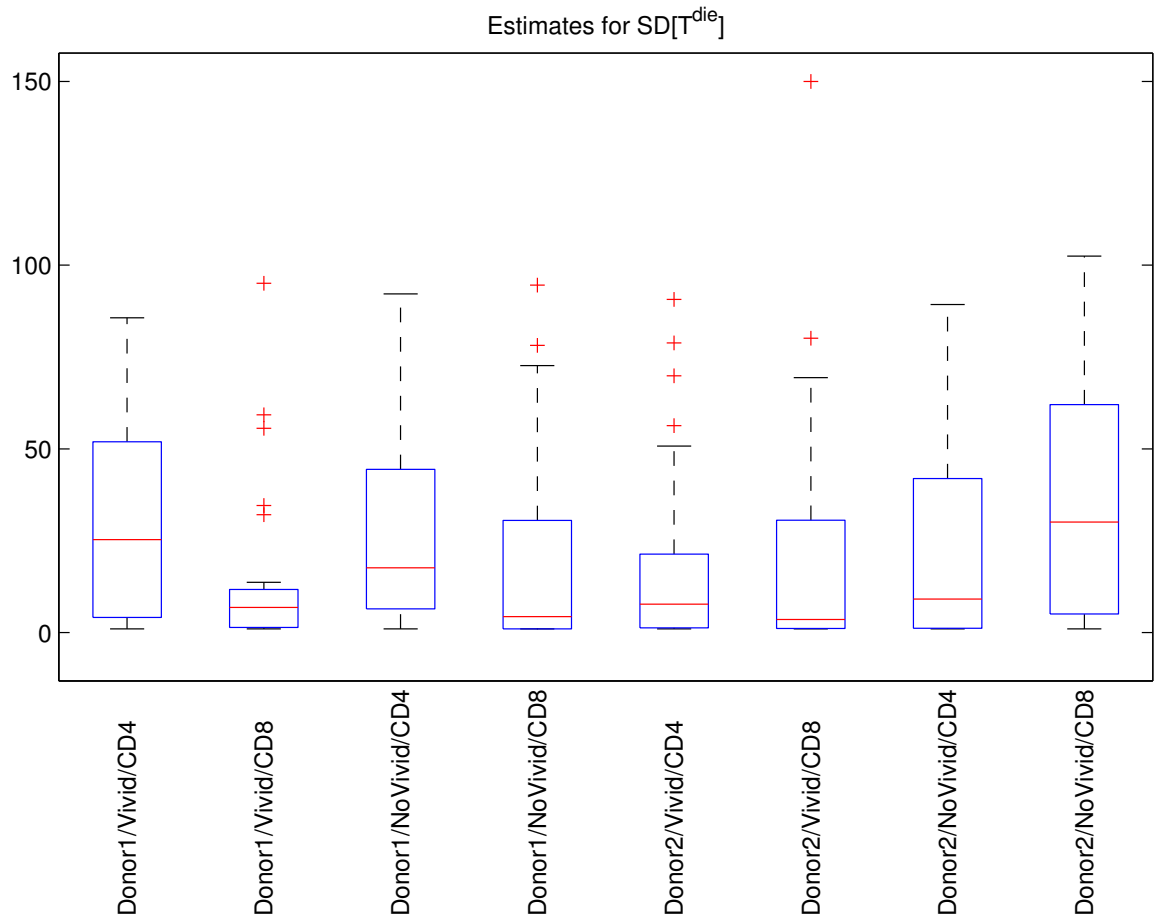


Figure 3.57: Box plots illustrating variability in estimates for the parameter SD $\left[T^{die}\right]$ when using only data from Days 1 through 3.

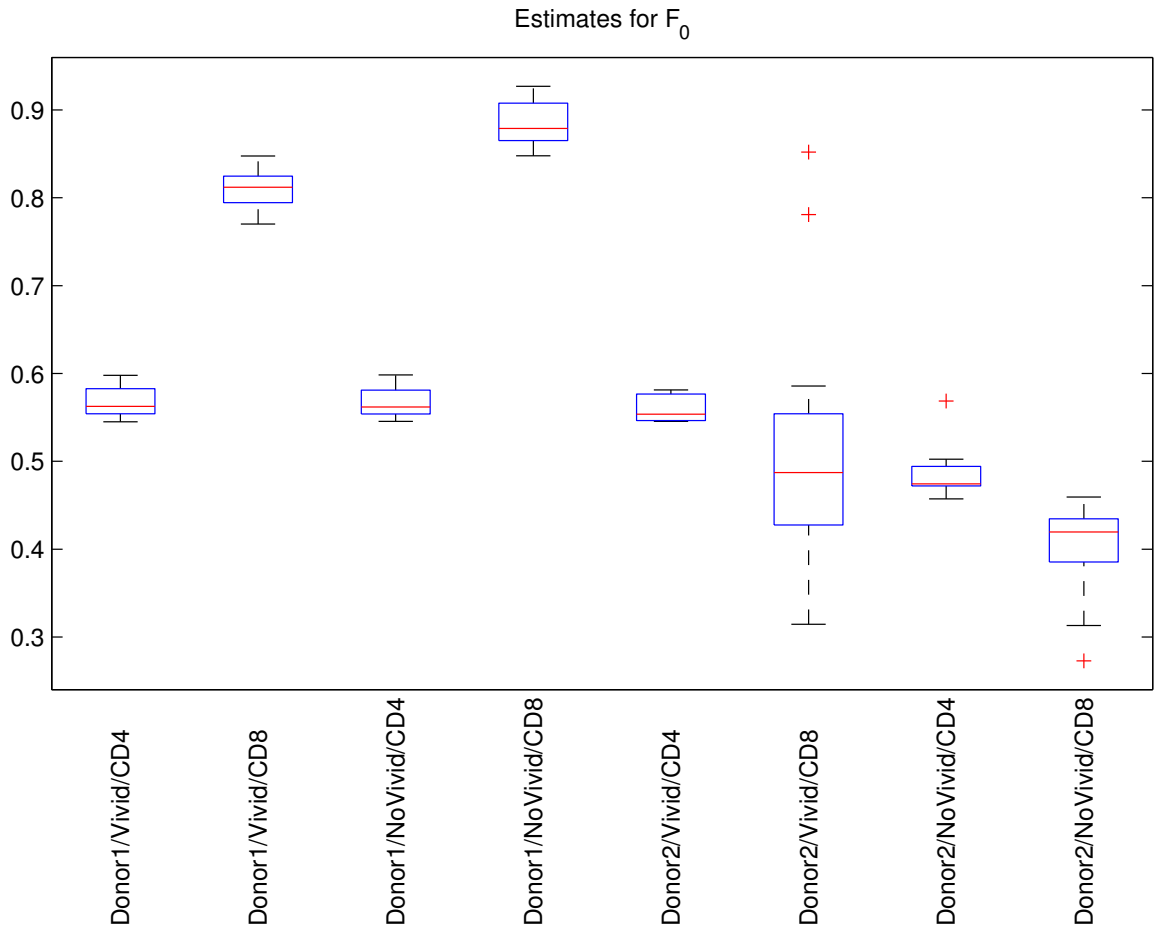


Figure 3.58: Box plots illustrating variability in estimates for the parameter F_0 when using only data from Days 1 through 3.

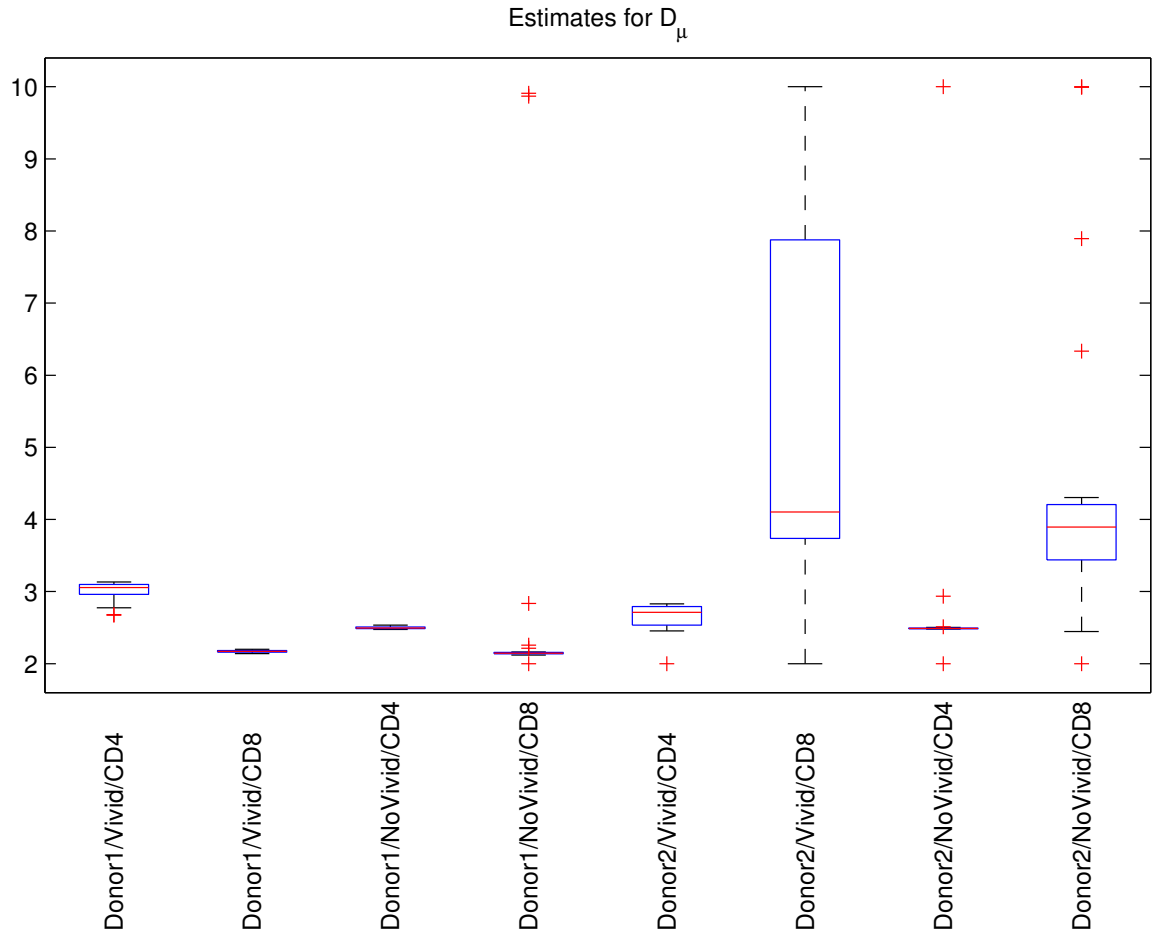


Figure 3.59: Box plots illustrating variability in estimates for the parameter D_μ when using only data from Days 1 through 3.

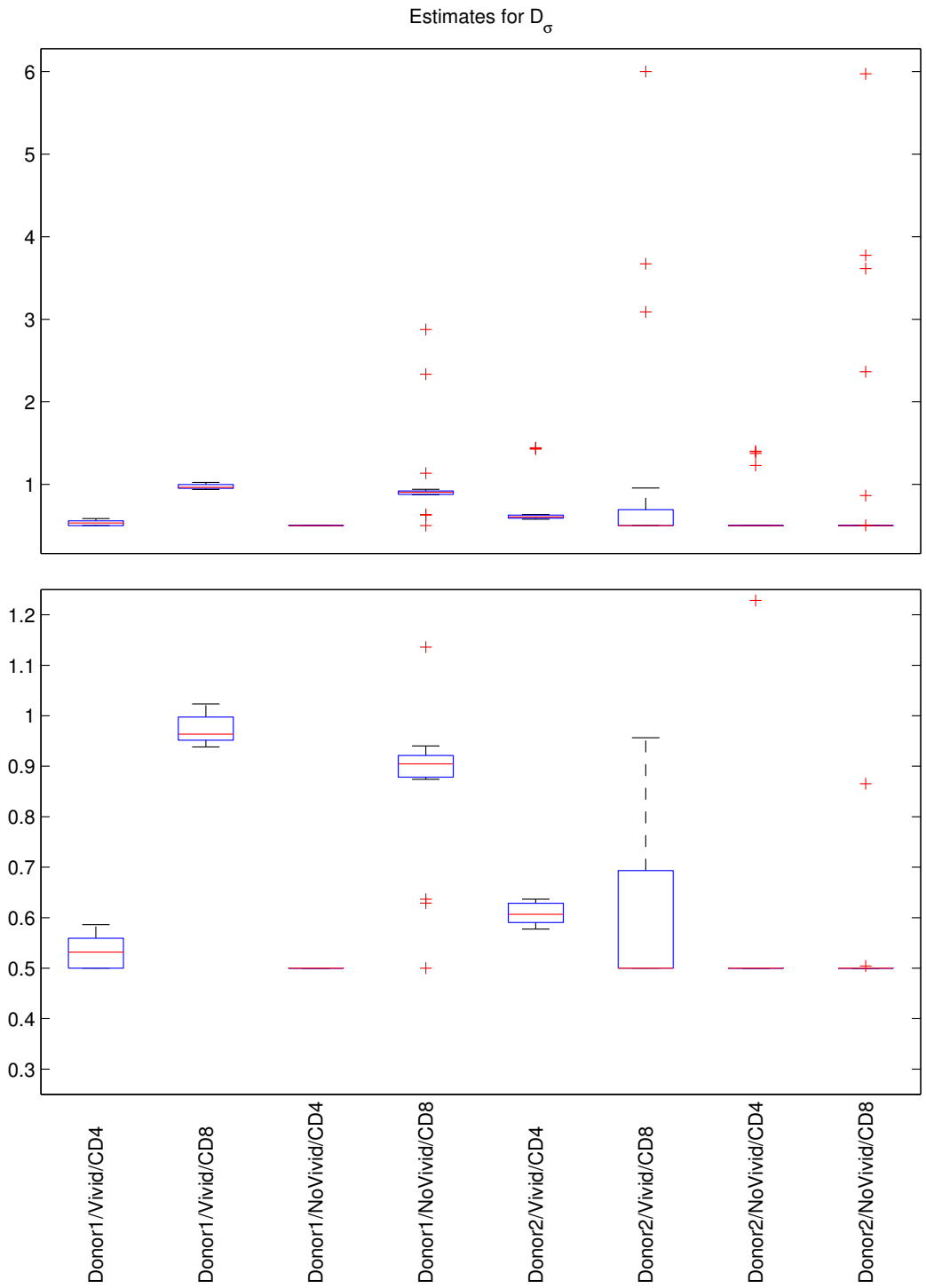


Figure 3.60: Box plots illustrating variability in estimates for the parameter D_σ when using only data from Days 1 through 3. In the lower set of box plots, some of the most extreme outliers are not plotted so that the rest of the information in the box plots can be shown with higher precision.

Comparing Figures 3.49 through 3.60 with Figures 3.10 through 3.21 reveals that there are clear differences between the parameter estimates obtained using three-day data sets and those obtained using complete data sets. For a complete discussion of the differences, see Section 5.2.5 of Banks et al. [4]. This outcome further supports the evidence given in Section 3.2 that the periodic exchange of nutrient medium starting at Day 3 causes changes to the counts and/or the proliferative behavior (as described by our cyton-based dynamical models) of the affected cell cultures.

3.4 Conclusions from Variability Study

In this chapter, we have presented the findings of our investigation into the variability that exists in CFSE-based flow cytometry data in the context of cyton-based mathematical models for cell proliferation. By applying the parameter estimation scheme described in Chapter 2 to a large body of data, we were able to assess both experimental and biological variability in the resulting parameter estimates. In this section, we summarize our findings and utilize them to make some important conclusions concerning both standard CFSE flow cytometry experimental procedures and the mathematical models that are used to analyze data that result from these procedures.

We begin by summarizing our results concerning identifiability of the various parameters in our model. As was discussed in Section 3.3.1, it appears that the parameters $\mathrm{E}\left[X_a\right]$, $\mathrm{SD}\left[X_a\right]$, c, $\mathrm{E}\left[T_0^{div}\right]$, $\mathrm{SD}\left[T_0^{div}\right]$, $\mathrm{E}\left[T^{div}\right]$, F_0, D_μ, and D_σ can all be estimated with fairly high reliability, while the parameters $\mathrm{SD}\left[T^{div}\right]$, $\mathrm{E}\left[T^{die}\right]$, and $\mathrm{SD}\left[T^{die}\right]$ do not appear to be identifiable. As a possible explanation of this, we proposed that the model may not be sensitive to the parameters involving T^{die} at the earlier time points (i.e., Days 1 through 3). We also provided an example (cf. Figures 3.22 through 3.24) which indicated that the model may be more sensitive to these parameters at the later time points (i.e., by Day 5) and used statistically based model comparison tests to verify this hypothesis. Unfortunately, another key finding of this study is that the accuracy of *in vitro* CFSE flow cytometry data appears to decrease significantly after Day 3. We hypothesize that cell culturing protocols (and specifically the depletion and replenishment of nutrient medium) lead to this increase of variability in measured triplicate data after Day 3. Thus the parameters $\mathrm{E}\left[T^{die}\right]$ and $\mathrm{SD}\left[T^{die}\right]$ may not be identifiable using five-day time series data sets collected using the current standard protocols.

A number of differences between the parameter estimates observed for distinct donors and cell types are outlined in Section 3.3.1, but it is difficult to state the relationships described there in a more concise way. We would like to revisit, however, one interesting general result that arose in Section 3.3.1: the use of ViViD dye does not seem to have a significant effect on the estimates obtained for most of the model parameters. Interestingly, the largest fractions of dead cells are

typically identified (using ViViD) in the first three days of the experiment (cf. Tables 3.1 through 3.5); i.e., the largest relative errors in cell counts due to the counting of dead cells (when not using ViViD to omit them) occurs in the earliest days of the experiment. This may be because the large die-off of cells that occurs immediately following PHA stimulation (which was mentioned in Section 2.5) leaves a considerable number of dead (but not yet disintegrated) cells in its wake at the beginning of the experiment. Presumably these cells are able to disintegrate completely by Day 4, at which point the fractions of dead cells identified using ViViD are much smaller. Whatever the reason, it seems that errors in cell counts caused by the inclusion of dead cells are likely to be largest when errors in the true cell counts, themselves, are smallest (cf. Section 3.2). Also, errors in cell counts caused by the inclusion of these cells are likely to be smallest when errors in the true cell counts, themselves, are largest. So, although the use of ViViD dye may not appear to be beneficial in the estimation of parameters based on the results of this study, use of ViViD may prove to be beneficial in later studies if new techniques make it possible to obtain more precise true cell counts (especially at the later time points).

We implemented several variations to our basic parameter estimation scheme in an attempt to reduce variability in parameter estimates. In Section 3.3.3, we discussed the results of fixing the values of $\mathrm{E}\left[T_0^{div}\right]$ or $\mathrm{E}\left[T^{die}\right]$, and concluded that neither of these schemes proved to be universally beneficial. In fact, in many cases these approaches tended to *increase* variability in parameter estimates. While Figures 13 and 14 indicate that we can (in some instances) expect a correlation between $\mathrm{E}\left[T_0^{div}\right]$ and $\mathrm{SD}\left[T_0^{div}\right]$ and between $\mathrm{E}\left[T^{die}\right]$ and $\mathrm{SD}\left[T^{die}\right]$, they do not necessarily imply that $\mathrm{E}\left[T_0^{div}\right]$ and/or $\mathrm{E}\left[T^{die}\right]$ are not influential. In fact, there may be some interplay between these parameters and other model parameters similar to the hypothetical interplay between time-to-division and time-to-death parameters discussed at the end of Section 3.3.1. Furthermore, even if the parameters in question are not influential, fixing them will not necessarily result in improved estimates for other parameters. (The data itself may contain little or no information on these specific parameters.) Thus, the discovery that fixing one or both of the parameters in question was not advantageous should not be particularly surprising or counterintuitive.

We also implemented a variation of our basic parameter estimation scheme in which only data from Days 1 through 3 were utilized. As was outlined in Section 3.3.4, clear differences exist between the parameter estimates obtained using three-day data sets and those obtained using complete data sets. This outcome, along with the evidence presented in Section 3.2, seems to indicate that the periodic exchange of nutrient medium starting at Day 3 does affect the proliferation of cells in the cell culture wells that are processed after Day 3. Furthermore, the very fact that the exchange of nutrient medium is necessary allows one to infer that cell cultures deplete nutrients in their respective wells throughout the experiment. Thus, several observations made in this study indicate that *the standard cell culturing protocol used for CFSE flow*

cytometry experiments does not provide a constant environment for cells growing in the wells. The apparent changes in the cell culture environments unfortunately contradict the underlying assumptions of our model. In particular, the cyton model that we've incorporated into our mass- and energy-conserving mathematical model tacitly assumes a *constant environment* for cultures of proliferating cells. (Note that the time-dependence of the cytons $\{(\phi_i(t), \psi_i(t))\}$ is actually based on "time since the last division occurred", or time in the frame of reference of an individual cell. The cytons therefore have no dependence on "real time", or time in the frame of reference of the experimenter. In real time, the conditions in the nutrient medium are apparently changing, and the basic cyton model is not capable of handling such a non-constant environment.) Adjustments could be made to the model, of course, to allow for time-dependence of the cytons, but any such adjustments would most likely lead to overparameterization; i.e., the increase in the number of parameters required by such adjustments would probably lead to even more identifiability issues. Furthermore, environmental changes caused by the exchange of the nutrient medium at discrete time points would probably be very difficult (experimentally) to quantify. It is worth noting that, in cases where a smaller total number of cells is stimulated (e.g., in response to antigen-specific challenges) the depletion of nutrient medium may not be significant over the course of a five-day experiment; thus, the exchange of nutrient medium may be unnecessary in such cases.

As a final point, we would like to re-emphasize that typical CFSE flow cytometry data *are not truly longitudinal* (cf. Section 3.1). For this reason, it appears that our ability to validate (and estimate parameters for) our division- and label-structured cell population model is limited by the precision with which experimenters can seed a set of cell culture wells. (Recall that, upon seeding, we assume all wells to be identical in that they include the same numbers of total cells in the same proportions.) The results presented in this report suggest that the use of time series data which are not truly longitudinal (which is the current standard protocol for *in vitro* CFSE flow cytometry experiments) leads to high variability in estimates for many of the relevant parameters in our model. This could be because the numbers of cells used to seed two distinct wells, which are assumed to be identical and which will be harvested and analyzed to produce data for two distinct points in the time series, can differ by 16 percent or more (cf. Table 3.1). (We should mention here the possibility that experimenters may actually seed the wells with much higher precision than seems to be the case, and that poor precision in *bead counts* might be the true cause of the apparent discrepancies in total cell numbers. This possibility was considered in our previous work [5], and a statistical model was formulated there to account for errors in bead counts, but we believe such a model adjustment leads to overparameterization.) This issue of poor precision in total cell counts, along with the issue of non-constant environment, suggests that substantive changes to CFSE flow cytometry and *in vitro* cell culturing protocol and/or significant modifications to our mathematical model need

to be considered.

We reiterate that, for this study, we have only considered PHA-stimulated cells. Because it is a *non-specific* T cell mitogen, PHA stimulates *all* T cells to begin dividing and therefore leads to a somewhat artificial situation from an immunological perspective. Nevertheless, the methods and results presented here indicate that many of the important and biologically relevant parameters for describing T cell proliferation can be reliably estimated using our approach. As more realistic and interesting experiments are devised and carried out (e.g., Gag protein stimulation of cells from HIV-positive donors), it is our hope that the challenges and concerns we have discussed will inform the development and selection of models that account for variability in experimental data.

Chapter 4

Computation of Relevant Convolution Integrals

For the cell proliferation models described in Chapter 2, we saw that the total fluorescence intensity (FI) emitted by any given cell in the range of wavelengths corresponding to CFSE is actually the sum of the cell's CFSE-induced FI and its natural autofluorescence. Therefore, if we are interested in the distribution of the total CFSE-wavelength FI for a population of cells, we actually need to compute the distribution of a random variable $\tilde{X}$ that is the sum of two independent random variables X and X_a that represent CFSE-induced FI and autofluorescence, respectively. One approach to this problem is to use the standard convolution formula [18], to obtain the probability density function (pdf) of $\tilde{X} = X + X_a$ from the (known) pdfs of X and X_a. That is, one can determine the value the pdf of $\tilde{X}$ at $\tilde{x}$ as

$$f_{\tilde{X}}(\tilde{x}) = \int_{-\infty}^{\infty} f_X(\xi) f_{X_a}(\tilde{x} - \xi)\, d\xi = \int_{-\infty}^{\infty} f_{X_a}(\xi) f_X(\tilde{x} - \xi)\, d\xi, \tag{4.1}$$

where f_X and f_{X_a} are the pdfs for CFSE-induced FI and autofluorescence, respectively, and $f_{\tilde{X}}$ is the pdf for total (or observed) FI. Unfortunately, this approach can be very computationally expensive, as it requires one to compute a new integral (over a potentially very large domain) for each value $\tilde{x}$ at which one wishes to know the density. In this chapter we examine methods for computing values for convolution integrals relevant to the cell proliferation models already described, as well as the asymmetric division models that will be described in Chapter 5.

4.1 Direct Methods

First, we consider two methods for computing the value of the integral in (4.1) directly. In particular, we examine one method based on the composite trapezoid rule quadrature and

another based on Monte Carlo sampling. These are both "direct" methods in the sense that they converge to the exact value of the integral as larger numbers of quadrature or sample points are used.

Before we describe the direct methods in detail, we take a moment to make four important points. The first is that fluorescence intensity (FI) *always takes on a nonnegative value*. Thus, the support for both f_X and f_{X_a} is some subset of the interval $[0, \infty)$ and the domain of integration for the integral in (4.1) can be truncated as in

$$f_{\tilde{X}}(\tilde{x}) = \int_0^{\tilde{x}} f_X(\xi) f_{X_a}(\tilde{x} - \xi)\, d\xi. \tag{4.2}$$

Second, we generally need to compute a *collection* of N range values for $f_{\tilde{X}}$ corresponding to some set of domain values $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ in order to obtain an approximation for the entire distribution of observed CFSE FI. Any computational scheme should, of course, take this need into account rather than focusing on obtaining single range values for $f_{\tilde{X}}$ one at a time. Third, in order to perform the computations that follow, we need to determine a maximal domain value $\tilde{x}_{\max} = \tilde{x}_N$ at which $f_{\tilde{X}}$ has "significant" support. We define the "significant support region" of a pdf to be those domain values at which the pdf takes on values significantly larger than zero. The maximum observed CFSE FI for the data sets considered in this dissertation is around $10^{5.5}$ (cf. Figure 1.4), so we use the conservative value $\tilde{x}_{\max} = 10^6$ when working with our data. Finally, given the large spread of the domain values at which $f_{\tilde{X}}$ has support and the fact that CFSE FI are typically organized according to a logarithmic scale, we choose $\tilde{x}_1 = 10^0 = 1$ through $\tilde{x}_N = \tilde{x}_{\max} = 10^6$ to be N *logarithmically* equally spaced points on the interval $[1, 10^6]$. The set of domain values $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ can easily be assigned using the MATLAB function `logspace`. Note that the resulting domain values are therefore *not* equally spaced when considering them on a linear scale.

4.1.1 Trapezoid Rule Method

In order to compute a set of range values for $f_{\tilde{X}}$ using the composite trapezoid rule [40], we first define a set of nodes $\{\tilde{\xi}_0, \tilde{\xi}_1, \ldots, \tilde{\xi}_{N_{trap}}\}$ that partition the interval $[0, \tilde{x}_{\max}]$ into N_{trap} subintervals of *equal width*. This can be done by setting $\tilde{\xi}_k = k \cdot h$, where $h = \tilde{x}_{\max}/N_{trap}$ is the width of each subinterval. (Note that N_{trap} is not necessarily equal to N, and $\tilde{\xi}_k$ is not in general equal to $\tilde{x}_k$ because the former comes from a set of nodes which are equally spaced in the *linear* sense.) Then the composite trapezoid rule applied to (4.2) gives

$$f_{\tilde{X}}(\tilde{\xi}_j) \approx \begin{cases} 0 & \text{for } j = 0, \\ h \cdot T\Big(\big(f_X(\tilde{\xi}_0) f_{X_a}(\tilde{\xi}_j), f_X(\tilde{\xi}_1) f_{X_a}(\tilde{\xi}_{j-1}), \ldots, f_X(\tilde{\xi}_j) f_{X_a}(\tilde{\xi}_0)\big)\Big) & \text{for } j \geq 1, \end{cases}$$

where

$$T\big((y_0,\ldots,y_j)\big) = \frac{1}{2}\left(y_0 + 2\sum_{k=1}^{j-1} y_k + y_j\right). \tag{4.3}$$

Note that the function $T : \mathbb{R}^M \to \mathbb{R}$ is equivalent to the MATLAB function `trapz`, which takes as input a vector of length M (representing function values at nodes on a unit-spaced grid) and returns a scalar (representing the composite trapezoid rule integral approximation). Therefore, one could compute the range values $\{f_{\tilde{X}}^{trap}(\tilde{\xi}_j)\}$ by looping through the indices 1 through N_{trap} and applying `trapz`. This approach, which is outlined in Algorithm 4.1.1, tends to be relatively slow, however, so we use an alternate implementation based on the MATLAB function `conv`. The `conv` function takes as input two vectors u and v and "convolves" them to produce a new vector w. If $u = (u_0,\ldots,u_{N_{trap}})$ and $v = (v_0,\ldots,v_{N_{trap}})$, then $w = (w_0,\ldots,w_{2N_{trap}})$ where

$$w_j = \sum_{k=\max\{0,j-N_{trap}\}}^{\min\{j,N_{trap}\}} u_k v_{j-k}.$$

Thus, we obtain

$$\begin{aligned}
w_0 &= u_0v_0,\\
w_1 &= u_0v_1 + u_1v_0,\\
w_2 &= u_0v_2 + u_1v_1 + u_2v_2,\\
&\vdots\\
w_{N_{trap}} &= u_0v_{N_{trap}} + u_1v_{N_{trap}-1} + \cdots + u_{N_{trap}}v_0\\
&\vdots\\
w_{2N_{trap}} &= u_{N_{trap}}v_{N_{trap}}.
\end{aligned}$$

Now, suppose we let $u = \big(f_X(\tilde{\xi}_0),\ldots,f_X(\tilde{\xi}_{N_{trap}})\big)$ and $v = \big(f_{X_a}(\tilde{\xi}_0),\ldots,f_{X_a}(\tilde{\xi}_{N_{trap}})\big)$. Furthermore, assume $u_0 = f_X(\tilde{\xi}_0) = 0$ and $v_0 = f_{X_a}(\tilde{\xi}_0) = 0$. Then, if we pass $h\cdot u$ and v to the `conv` function, the output vector w will satisfy

$$\begin{aligned}
w_0 &= hf_X(\tilde{\xi}_0)f_{X_a}(\tilde{\xi}_0) = 0 = f_{\tilde{X}}^{trap}(\tilde{\xi}_0),\\
w_1 &= h\Big(f_X(\tilde{\xi}_0)f_{X_a}(\tilde{\xi}_1) + f_X(\tilde{\xi}_1)f_{X_a}(\tilde{\xi}_0)\Big) = h(0+0) = 0 = f_{\tilde{X}}^{trap}(\tilde{\xi}_1),\\
w_2 &= h\Big(f_X(\tilde{\xi}_0)f_{X_a}(\tilde{\xi}_2) + f_X(\tilde{\xi}_1)f_{X_a}(\tilde{\xi}_1) + f_X(\tilde{\xi}_2)f_{X_a}(\tilde{\xi}_0)\Big)\\
&\qquad = \frac{1}{2}h\Big(0 + 2f_X(\tilde{\xi}_1)f_{X_a}(\tilde{\xi}_1) + 0\Big) = f_{\tilde{X}}^{trap}(\tilde{\xi}_2),\\
&\vdots
\end{aligned}$$

$$w_{N_{trap}} = h\left(f_X(\tilde{\xi}_0) f_{X_a}(\tilde{\xi}_{N_{trap}}) + \sum_{k=1}^{N_{trap}-1} \left[f_X(\tilde{\xi}_k) f_{X_a}(\tilde{\xi}_{N_{trap}-k}) \right] + f_X(\tilde{\xi}_{N_{trap}}) f_{X_a}(\tilde{\xi}_0) \right)$$
$$= \frac{1}{2} h\Big(0 + 2 \sum_{k=1}^{N_{trap}-1} \left[f_X(\tilde{\xi}_k) f_{X_a}(\tilde{\xi}_{N_{trap}-k}) \right] + 0 \Big) = f_{\tilde{X}}^{trap}(\tilde{\xi}_{N_{trap}}),$$

where $f_{\tilde{X}}^{trap}(\tilde{\xi}_j)$ is the trapezoid rule computed value for $f_{\tilde{X}}(\tilde{\xi}_j)$ as given in Algorithm 4.1.1. We remark that the assumptions $f_X(\tilde{\xi}_0) = 0$ and $f_{X_a}(\tilde{\xi}_0) = 0$ (where $\tilde{\xi}_0 = 0$) are not unreasonable, as X_a typically has a lognormal distribution and X can usually be approximated well by a sum of lognormal random variables. Algorithm 4.1.2, which utilizes the `conv` function, produces results identical to Algorithm 4.1.1 under these assumptions and requires substantially less computational time.

Algorithm 4.1.1 Trapezoid Rule Convolution Computation Using `trapz`

1. Define $N_{trap} + 1$ (linearly) equally spaced nodes $\{\tilde{\xi}_0, \ldots, \tilde{\xi}_{N_{trap}}\}$ with spacing h.
2. Set $f_{\tilde{X}}^{trap}(\tilde{\xi}_0)$ to 0.
3. For each j in $\{1, \ldots, N_{trap}\}$, do the following:
   - Construct the vector $u = \big(f_X(\tilde{\xi}_0), \ldots, f_X(\tilde{\xi}_j)\big)$.
   - Construct the vector $v = \big(f_{X_a}(\tilde{\xi}_j), \ldots, f_{X_a}(\tilde{\xi}_0)\big)$, where the indices run in *decreasing* order.
   - Compute the vector $y = u. * v$, where ".*" indicates element-wise multiplication.
   - Compute $T(y)$ (cf. (4.3)) using `trapz` and then set $f_{\tilde{X}}^{trap}(\tilde{\xi}_j)$ to $h \cdot T(y)$.
4. Apply piecewise linear interpolation to the ordered pairs in $\{(\tilde{\xi}_k, f_{\tilde{X}}^{trap}(\tilde{\xi}_k)) : k \in \{0, \ldots, N_{trap}\}\}$ to obtain the values $\{f_{\tilde{X}}(\tilde{x}_k) : k \in \{1, \ldots, N\}\}$. This can be effected using the MATLAB function `interp1`.

Algorithm 4.1.2 Trapezoid Rule Convolution Computation Using `conv`

1. Define $N_{trap}+1$ (linearly) equally spaced nodes $\{\tilde{\xi}_0,\ldots,\tilde{\xi}_{N_{trap}}\}$ with spacing h.
2. Construct the vector $h\cdot u=\big(hf_X(\tilde{\xi}_0),\ldots,hf_X(\tilde{\xi}_{N_{trap}})\big)$.
3. Construct the vector $v=\big(f_{X_a}(\tilde{\xi}_0),\ldots,f_{X_a}(\tilde{\xi}_{N_{trap}})\big)$.
4. Compute the vector w by convolving the the vectors $h\cdot u$ and v using `conv`.
5. For each j in $\{0,\ldots,N_{trap}\}$, set $f_{\tilde{X}}^{trap}(\tilde{\xi}_j)$ to w_j. (Note that elements $N_{trap}+1$ through $2N_{trap}$ of the vector w are never used.)
6. Apply piecewise linear interpolation to the ordered pairs in $\{(\tilde{\xi}_k,f_{\tilde{X}}^{trap}(\tilde{\xi}_k)) : k\in\{0,\ldots,N_{trap}\}\}$ to obtain the values $\{f_{\tilde{X}}(\tilde{x}_k) : k\in\{1,\ldots,N\}\}$. This can be effected using the MATLAB function `interp1`.

4.1.2 Monte Carlo Method

The idea behind the Monte Carlo approach to computing the pdf $f_{\tilde{X}}$ is completely different. Given that $\tilde{X}=X+X_a$, we sample a collection of N_{mc} values $\{x^k\}$ from f_X and another collection of N_{mc} values $\{x_a^k\}$ from f_{X_a}. This allows us to construct a collection of N_{mc} values $\{\tilde{x}_{mc}^k\}$ by setting $\tilde{x}_{mc}^k=x^k+x_a^k$. The values in $\{\tilde{x}_{mc}^k\}$ form a realization of a random sample from the distribution $f_{\tilde{X}}$; i.e., these values represent N_{mc} realizations of $\tilde{X}$. Therefore, if we define a set of $N_{hist}\ll N_{mc}$ bins (or subintervals) on the interval $[0,\tilde{x}_{\max}]$, we can generate a relative frequency histogram [21] and approximate $f_{\tilde{X}}$ by the piecewise linear spline that connects the midpoints of the tops of the histogram bars. Kernel density estimation (KDE) provides a more sophisticated approach to the problem of approximating a probability density function using a discrete sample [16], and results in a smooth density approximation rather than a piecewise linear spline. We use KDE in the form of the MATLAB function `ksdensity` to generate an approximation of $f_{\tilde{X}}$ from the finite sample $\{\tilde{x}_{mc}^k\}$ in our Monte Carlo implementation. This implementation is outlined in Algorithm 4.1.3.

Algorithm 4.1.3 Monte Carlo Convolution Computation

1. Construct the vector $x \in \mathbb{R}^{N_{mc}}$ by randomly sampling from the distribution f_X.
2. Construct the vector $x_a \in \mathbb{R}^{N_{mc}}$ by randomly sampling from the distribution f_{X_a}.
3. Compute the vector $\tilde{x}_{mc}$ as the sum of the vectors x and x_a.
4. Obtain the vector $f_{\tilde{X}}^{mc}$, which represents the approximate values of $f_{\tilde{X}}$ at each of the N domain values in $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$, by applying `ksdensity` to the vector $\tilde{x}_{mc}$.

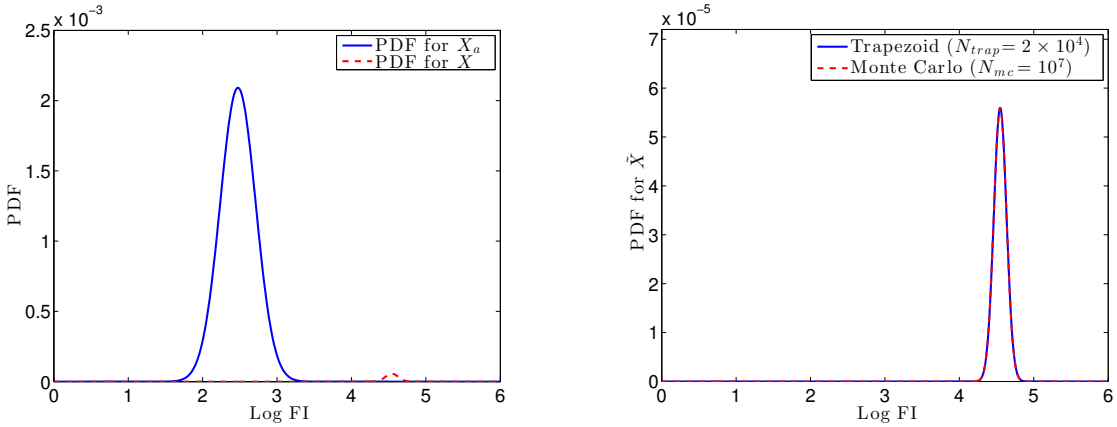
4.1.3 Testing the Direct Methods

In order to test and compare the two direct approaches we've described, we propose two test problems. In both of these, we suppose $X_a \sim \text{logn}(6, 0.55^2)$, which is a realistic assumption based on estimated autofluorescence parameters obtained in Chapter 3. Then, in the first test problem, we suppose $X \sim \text{logn}(10.5, 0.2^2)$, which is a realistic assumption for CFSE FI data at Day 1. In the second test problem, we suppose $X \sim \text{logn}(6.8, 0.8^2)$, which is a realistic assumption for CFSE FI data at Day 5. For both of the test problems, we estimate the values of $f_{\tilde{X}}$ at 1025 logarithmically evenly spaced nodes on the interval $[1, 10^6]$.

The results of applying our two direct methods to the first test problem are shown in Figure 4.1. As indicated in the figure, we set the number of trapezoid rule nodes to $N_{trap} = 2 \times 10^4$ (which corresponds to a step size $h = 50$) and the number of Monte Carlo sample points to $N_{mc} = 10^7$. The trapezoid rule pdf approximation for $\tilde{X}$, which was obtained using Algorithm 4.1.2, required 0.107 seconds to produce. The Monte Carlo pdf approximation, which was obtained using Algorithm 4.1.3, required 11.9 seconds to produce. (Note that all timings provided in Sections 4.1 and 4.2 are based upon runs using MATLAB Release 2013a on an early 2009 MacBook Pro with a 2.93 GHz Intel Core 2 Duo processor and 8 GB of 1067 MHz DDR3 memory.) The two pdf approximations are virtually indistinguishable in the figure, indicating excellent agreement.

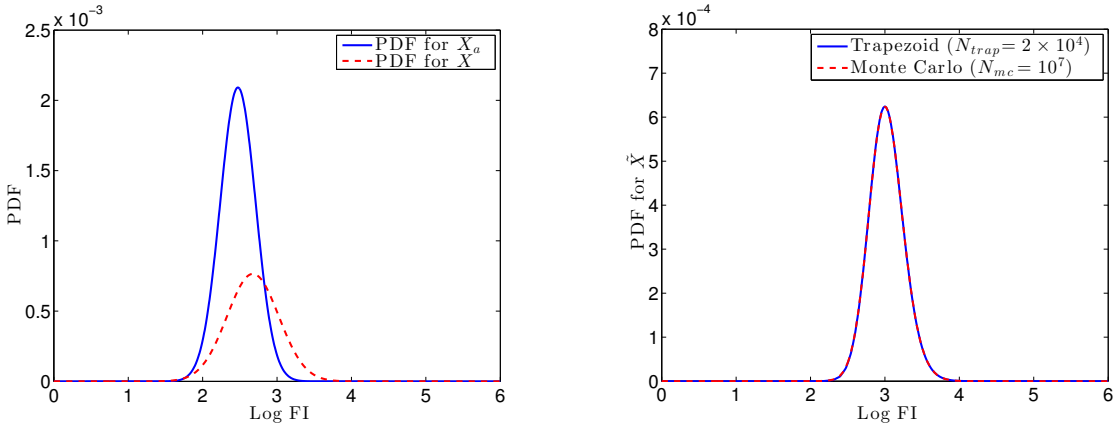
The results of applying our two direct methods to the second test problem are shown in Figure 4.2. Again, we set $N_{trap} = 2 \times 10^4$ and $N_{mc} = 10^7$. The trapezoid rule pdf approximation for $\tilde{X}$, which was obtained using Algorithm 4.1.2, required 0.108 seconds to produce. The Monte Carlo pdf approximation, which was obtained using Algorithm 4.1.3, required 20.8 seconds to produce. The two pdf approximations are virtually indistinguishable in the figure, once again indicating excellent agreement.

In order to demonstrate the convergence of the trapezoid rule based method, we applied



(a) (b)

Figure 4.1: Plots of (a) f_{X_a} and f_X and (b) $f_{\tilde{X}}$ for the first test problem of Section 4.1.3.



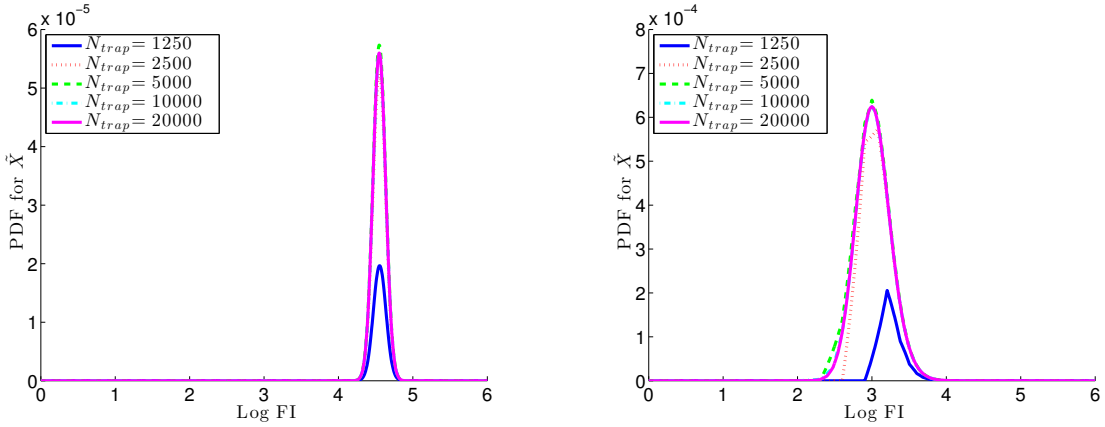
(a) (b)

Figure 4.2: Plots of (a) f_{X_a} and f_X and (b) $f_{\tilde{X}}$ for the second test problem of Section 4.1.3.

Algorithm 4.1.2 to the two test problems using increasing values for N_{trap}. The results are shown in Figure 4.3. We computed a "true" pdf for $\tilde{X}$ by applying Algorithm 4.1.2 with $N_{trap} = 10^5$, and used this pdf in order to compute errors for approximate pdfs shown in Figure 4.3. The maximum absolute errors are tabulated in Tables 4.1 and 4.2. The decreasing magnitude of the errors as N_{trap} increases indicates convergence of the algorithm.

Similarly, in order to demonstrate the convergence of the Monte Carlo method, we applied Algorithm 4.1.3 to the two test problems using increasing values for N_{mc}. The results are shown in Figure 4.4. Using the same "true" pdf described in the previous paragraph, we computed errors for approximate pdfs shown in Figure 4.4. The maximum absolute errors are tabulated in Tables 4.3 and 4.4. The decreasing magnitude of the errors as N_{mc} increases indicates that the Monte Carlo based algorithm also converges. Either algorithm can be used to obtain the desired pdf approximation to arbitrary precision, but comparing Tables 4.1 and 4.2 with Tables 4.3 and 4.4, we see that the trapezoid rule convolution method requires considerably less time than the Monte Carlo method. More specifically, to achieve a level of precision such that the maximum absolute errors that are less than about 1% of the maximum value of $f_{\tilde{X}}$, the trapezoid rule convolution method requires about 0.10 seconds while the Monte Carlo convolution method requires on average about 150 seconds. Thus, the trapezoid rule method appears to be the more efficient of the two direct methods.

Before moving on to discuss indirect approximation methods for the convolution formula, we would like to make one more remark concerning our two test problems and the trapezoid rule method. The first test problem is apparently "easier" to solve than the second in the sense that fewer trapezoid rule quadrature points are required to attain a comparable level of accuracy. This can be explained by the fact that in the first test problem the most significant region of support for f_X consists of values that are considerably larger than those lying in the most significant region of support for f_{X_a} (cf. Figure 4.1(a)). Thus, when adding the random variables X and X_a, the latter tends to contribute relatively little to the sum. Also, because the scale for the significant support of the former tends to be so large, a large step size h (and a correspondingly small N_{trap}) is generally sufficient for approximating the pdf of X and thus of $X + X_a$. The first test problem is a special case of the general situation encountered in the early days of a CFSE-base flow cytometry experiment. During these early days, autofluorescence makes a negligible contribution to total FI, but as the experiment progresses autofluorescence becomes more important. Therefore, it is important to consider the (temporal) length of the experiments when choosing N_{trap}.



(a) (b)

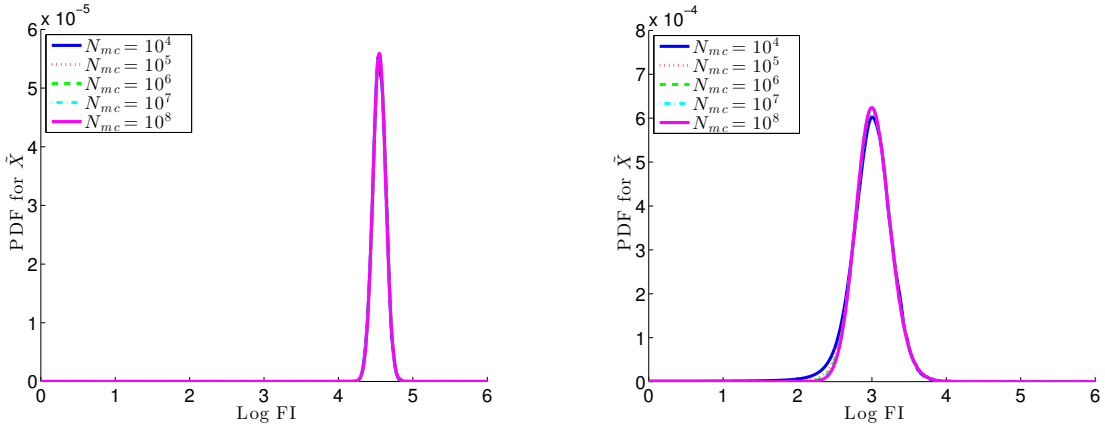
Figure 4.3: Plots of trapezoid rule approximations of $f_{\tilde{X}}$ showing convergence for (a) the first test problem and (b) the second test problem of Section 4.1.3.

Table 4.1: Convergence of Trapezoid Rule Convolution Method (Algorithm 4.1.2) for the first test problem of Section 4.1.3.

N_{trap}	h	Time (s)	Error
1250	800	0.0018998	3.6352×10^{-5}
2500	400	0.0022264	3.5371×10^{-6}
5000	200	0.0095202	1.3668×10^{-6}
10000	100	0.0285389	2.7076×10^{-8}
20000	50	0.0929019	9.7830×10^{-10}

Table 4.2: Convergence of Trapezoid Rule Convolution Method (Algorithm 4.1.2) for the second test problem of Section 4.1.3.

N_{trap}	h	Time (s)	Error
1250	800	0.0018564	5.8581×10^{-4}
2500	400	0.0022326	1.1111×10^{-4}
5000	200	0.0104667	4.0703×10^{-5}
10000	100	0.0303276	7.5597×10^{-6}
20000	50	0.0960450	1.7567×10^{-6}



(a) (b)

Figure 4.4: Plots of Monte Carlo approximations of $f_{\tilde{X}}$ showing convergence for (a) the first test problem and (b) the second test problem of Section 4.1.3.

Table 4.3: Convergence of Monte Carlo Convolution Method (Algorithm 4.1.3) for the first test problem of Section 4.1.3.

N_{mc}	Time (s)	Error
10^4	0.19577	2.70889×10^{-6}
10^5	0.19024	1.38063×10^{-6}
10^6	1.19868	4.00708×10^{-7}
10^7	11.2137	1.29064×10^{-7}
10^8	123.754	5.48362×10^{-8}

Table 4.4: Convergence of Monte Carlo Convolution Method (Algorithm 4.1.3) for the second test problem of Section 4.1.3.

N_{mc}	Time (s)	Error
10^4	0.22124	3.73811×10^{-5}
10^5	0.28094	1.65752×10^{-5}
10^6	2.29956	7.64454×10^{-6}
10^7	19.8908	2.94373×10^{-6}
10^8	184.280	1.25757×10^{-6}

4.2 Indirect Methods

In working with some families of distributions, the distribution of the sum of two random variables can be expressed exactly. For example, the sum of any two normally distributed random variables will have a normal distribution [18]. Unfortunately, in our application we typically need to find the distribution of the sum of two *lognormally* distributed random variables and we cannot make a similar claim for such a sum. It has been argued, however, that the sum of two lognormally distributed random variables has a distribution that is *approximately* lognormal [25, 42]. If one can easily determine values for the two parameters μ and σ^2 that define the approximating lognormal distribution, the approximate values of the pdf at a set of domain points $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ can be computed much more quickly than is possible through direct methods.

Finding the pdf for a sum of two or more lognormally distributed random variables is an important problem in telecommunications and therefore a variety of approaches to approximating such a pdf have been proposed by researchers in that field [13]. Two popular approximations have been proposed by Fenton [25] and Schwartz and Yeh [42]. Hereafter, we refer to these approximation techniques as "indirect" methods to distinguish them from the direct methods discussed in Section 4.1. In the Fenton method (which is sometimes referred to as the Wilkinson or Fenton-Wilkinson method) one approximates the pdf of the sum with a lognormal distribution that has the same mean and variance as the true distribution of the sum. In the Schwartz-Yeh method, on the other hand, one computes the mean and variance of $Z = \log(X + X_a)$ and uses these as the parameters for the approximating lognormal distribution. Both methods allow for the approximation of the pdf of a sum of two *or more* lognormally distributed random variables, but in our application we only ever need to consider the sum of two random variables.

4.2.1 Fenton Method

As asserted above, the Fenton method [25] determines the mean and variance of the distribution the *sum* of X and X_a (which are both assumed to be lognormal) and uses the lognormal distribution with this mean and variance as an approximation of the distribution of $\tilde{X} = X + X_a$. In order to outline the method, we will utilize the following facts concerning sums of random variables and properties of lognormal distributions [18]. First of all, the mean and variance of the sum of two independent random variables X and X_a are given by

$$\mathrm{E}\left[X + X_a\right] = \mathrm{E}\left[X\right] + \mathrm{E}\left[X_a\right] \tag{4.4}$$

and

$$\mathrm{Var}\left[X + X_a\right] = \mathrm{Var}\left[X\right] + \mathrm{Var}\left[X_a\right], \tag{4.5}$$

respectively. Also, the mean and variance of a random variable Y that is lognormally distributed with parameters μ and σ^2 (i.e., $Y \sim \text{logn}(\mu, \sigma^2)$) are

$$\text{E}[Y] = e^{\mu+(\sigma^2/2)} \tag{4.6}$$

and

$$\text{Var}[Y] = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2} = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}, \tag{4.7}$$

respectively. Alternatively, if Y is lognormally distributed and $\text{E}[Y]$ and $\text{Var}[Y]$ are known, one can solve the system of equations given by (4.6) and (4.7) to obtain expressions for the distribution parameters μ and σ. These are

$$\mu = \log\left(\text{E}[Y]\right) - \frac{1}{2}\log\left(1 + \frac{\text{Var}[Y]}{(\text{E}[Y])^2}\right) \tag{4.8}$$

and

$$\sigma^2 = \log\left(1 + \frac{\text{Var}[Y]}{(\text{E}[Y])^2}\right). \tag{4.9}$$

Algorithm 4.2.1, which makes use of the above formulae, outlines the Fenton method.

Algorithm 4.2.1 Fenton Method Convolution Approximation

---

1. Given μ_X and σ_X^2, which are the lognormal distribution parameters for X, compute $\text{E}[X]$ and $\text{Var}[X]$ using (4.6) and (4.7).

2. Given μ_{X_a} and $\sigma_{X_a}^2$, which are the lognormal distribution parameters for X_a, compute $\text{E}[X_a]$ and $\text{Var}[X_a]$ using (4.6) and (4.7).

3. Compute $\text{E}\left[\tilde{X}\right]$ and $\text{Var}\left[\tilde{X}\right]$, which are the mean and variance for $\tilde{X} = X + X_a$, using (4.4) and (4.5) along with the computed values $\text{E}[X]$, $\text{E}[X_a]$, $\text{Var}[X]$, and $\text{Var}[X_a]$.

4. Compute $\mu_{\tilde{X}}$ and $\sigma_{\tilde{X}}$ using (4.8) and (4.9) along with the computed values $\text{E}\left[\tilde{X}\right]$ and $\text{Var}\left[\tilde{X}\right]$.

5. Use the approximation

$$f_{\tilde{X}}(\tilde{x}) \approx \text{logn}(\tilde{x}; \mu_{\tilde{X}}, \sigma_{\tilde{X}}^2) = \frac{1}{\tilde{x}\sigma_{\tilde{X}}\sqrt{2\pi}} \cdot \exp\left[\frac{-(\log\tilde{x} - \mu_{\tilde{X}})^2}{2\sigma_{\tilde{X}}^2}\right]$$

   to obtain the values $\{f_{\tilde{X}}(\tilde{x}_k) : k \in \{1, \ldots, N\}\}$.

4.2.2 Schwartz-Yeh Method

Like the Fenton method, the Schwartz-Yeh method [42] is based on moment matching. Since X and X_a have lognormal distributions (by assumption), they can be expressed as $X = \exp(Y_1)$ and $X_a = \exp(Y_2)$, where Y_1 and Y_2 are Gaussian random variables. If we let

$$Z = \log\left(e^{Y_1} + e^{Y_2}\right),$$

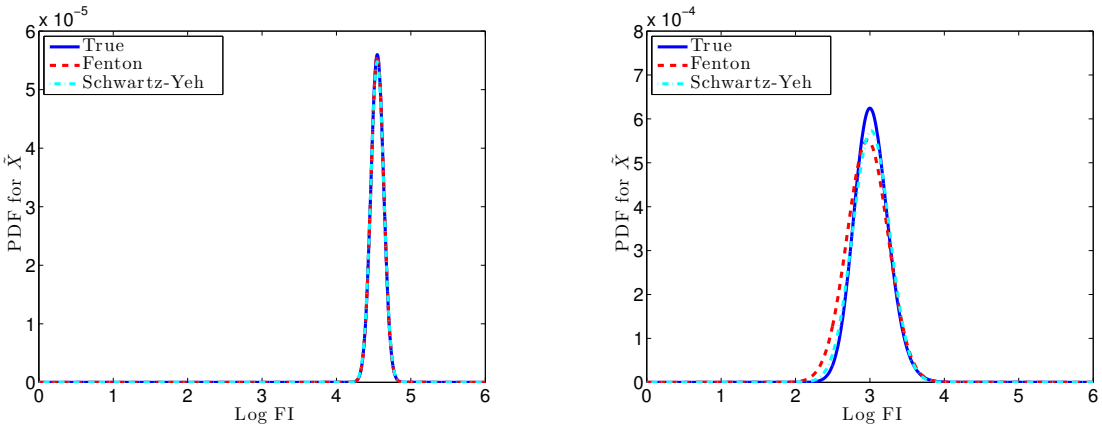
then

$$e^Z = e^{Y_1} + e^{Y_2} = X + X_a = \tilde{X}.$$

Thus, if $\tilde{X}$ is to be approximated by a lognormal distribution, the mean and variance of Z should be good candidates for the lognormal distribution parameters μ and σ^2. Because the computations involved in determining the mean and variance of Z are rather complicated, we utilize Takaki's MATLAB implementation [46] of the Schwartz-Yeh method rather than implementing the method ourselves. Note that Takaki's code uses a computationally efficient approach to obtaining the Schwartz-Yeh approximation as suggested by Ho [30].

4.2.3 Testing the Indirect Methods

To test and compare the Fenton and Schwartz-Yeh methods, we first apply them to the two test problems described in Section 4.1.3. We once again determine a "true" pdf for $\tilde{X}$ by applying Algorithm 4.1.2 with $N_{trap} = 10^5$ and use this for purposes of comparison and error computing. The results, which are shown in Figure 4.5, indicate that both approximation methods perform very well for the first test problem, but only reasonably well for the second. The timings and maximum absolute errors are tabulated in Table 4.5. There we see that the Fenton method performs better than the Schwartz-Yeh method, both in terms of computational time and accuracy, when applied to the first test problem. For the second test problem, the Schwartz-Yeh method produces a more accurate approximation, but again requires almost 100 times more computational time.



(a) (b)

Figure 4.5: Plots of Fenton and Schwartz-Yeh approximations of $f_{\tilde{X}}$ for (a) the first test problem and (b) the second test problem of Section 4.1.3.

Table 4.5: Timing and maximum absolute errors for Fenton and Schwartz-Yeh approximations as applied to the first (1st) and second (2nd) test problems of Section 4.1.3.

Test Problem	Method	Time (s)	Error
1st	Fenton	0.0004418	9.8608×10^{-8}
	Schwartz-Yeh	0.0174341	1.2081×10^{-6}
2nd	Fenton	0.0004875	1.1775×10^{-4}
	Schwartz-Yeh	0.0365148	5.6341×10^{-5}

4.3 Fenton Method vs. Trapezoid Method

As discussed in Section B.4 of the Appendix, we use the Fenton method to compute convolution integrals when working with the symmetric division cell proliferation models described in Chapter 2. While the Fenton method is not as accurate as the trapezoid rule convolution method, or even the Schwartz-Yeh method in some situations, it requires considerably less computational time than either of these. Because of the computational speed allowed by the Fenton method, forward solutions of the mathematical model specified in Section 2.5 require only about 0.13 seconds and it was possible to solve the inverse (parameter estimation) problems described in Chapter 3 in an average of 6.5 minutes per data set. (Timings provided in this section are based upon runs using MATLAB Release 2012a on a Dell Optiplex 990 with eight (8) 3.4 GHz Intel Core i7-2600 processors and 8 GB of 1333 MHz memory.) In contrast, forward solutions of the same mathematical model require about 21.5 seconds when using the trapezoid rule convolution method.

In this section, we consider how the decision to use the less accurate Fenton method over the trapezoid rule convolution method might have affected the results discussed in Chapter 3. Because of the computational time required, we choose to perform new parameter estimations only for a select subset of the data sets that were considered in that chapter. Recall that in Section 3.3 we considered $4 \times 162 = 648$ data sets for Donor 1 and $4 \times 243 = 972$ data sets for Donor 2. In order to compare the parameter estimates obtained there with a new collection of parameter estimates obtained using the trapezoid rule convolution method, we identify the three lowest cost and three highest cost data sets for each donor and cell type, removing from consideration all data sets for which ViViD dye was *not* used. Note that the "cost" of a data set is determined by the value of the GLS cost functional given in (2.17). The 24 data sets satisfying these criteria are listed in Table 4.6. Note also that we use the same abbreviations to denote experimental combinations of donor, ViViD dye status, and cell type as were used in Section 3.3, and the same form for the Data IDs as was used in Table 3.19 of that section.

Now, we apply our parameter estimation procedure to the selected data sets as was done in Section 3.3, except that we utilize the trapezoid rule convolution method instead of the Fenton method. To be more specific, we compute each convolution integral in the last line of (B.9) (in Section B.4.1 of the Appendix) using Algorithm 4.1.2. To compare the results, we list the percent change in the costs and the absolute percent change in the parameter estimates for each of the selected data sets in Table 4.7. In this table, each percent change is calculated as

$$\frac{v_{tm} - v_{fm}}{v_{fm}} \times 100\%,$$

where v_{tm} and v_{fm} denote the values obtained when using the trapezoid rule based method

Table 4.6: Data sets selected for comparing parameter estimates obtained using Fenton method with those obtained using trapezoid rule convolution method.

Data Set	Experiment	Data ID	Cost
1	Donor1/Vivid/CD4	10012	12415.5
2	Donor1/Vivid/CD4	10011	12642.6
3	Donor1/Vivid/CD4	11012	12710.0
4	Donor1/Vivid/CD4	21211	32661.5
5	Donor1/Vivid/CD4	11211	32698.6
6	Donor1/Vivid/CD4	11111	33047.9
7	Donor1/Vivid/CD8	21111	8446.42
8	Donor1/Vivid/CD8	20111	8451.07
9	Donor1/Vivid/CD8	10111	8774.17
10	Donor1/Vivid/CD8	01200	21334.6
11	Donor1/Vivid/CD8	02000	21607.9
12	Donor1/Vivid/CD8	02200	22621.2
12	Donor2/Vivid/CD4	00222	10210.0
14	Donor2/Vivid/CD4	10222	10214.6
15	Donor2/Vivid/CD4	00212	10526.7
16	Donor2/Vivid/CD4	00100	16099.4
17	Donor2/Vivid/CD4	11110	16290.7
18	Donor2/Vivid/CD4	01100	16435.0
19	Donor2/Vivid/CD8	02001	10256.4
20	Donor2/Vivid/CD8	00002	10419.4
21	Donor2/Vivid/CD8	02020	10471.5
22	Donor2/Vivid/CD8	20211	15565.6
23	Donor2/Vivid/CD8	21211	16008.1
24	Donor2/Vivid/CD8	21111	16099.3

and Fenton method, respectively, and each *absolute* percent change is calculated as the absolute value of this quantity.

In Table 4.7, we clearly see that the results obtained when using the trapezoid rule convolution method (TM) differ from those obtained with the Fenton method (FM). For 17 of the 24 data sets considered, the GLS cost is lower when using TM than FM. On average, the cost associated with TM is 11.68% less than that associated with FM, and median relative change in cost when moving from FM to TM is a 6.53% decrease. Thus, the costs tend to indicate that better model fits can be attained with TM than FM. The differences observed in the parameter estimates, themselves, cause us greater concern. Only 5 of the 12 parameters experience a median absolute percent change less than 5% when moving from FM to TM, and 6 of the 12 parameters show an absolute percent change of greater than 10%. On a positive note, the 5 parameters showing the smallest median absolute percent change include 4 of the parame-

ters that are most important in describing cell proliferation: E $\left[T_0^{div}\right]$, SD $\left[T_0^{div}\right]$, E $\left[T^{div}\right]$, and F_0. Three other important parameters, SD $\left[T^{div}\right]$, E $\left[T^{die}\right]$, and SD $\left[T^{die}\right]$, all experienced very large median absolute percent change when moving from FM to TM, but these parameters are unlikely to be identifiable when using our data for reasons described in detail Chapter 3.

The results presented in Table 4.7 suggest that it might be worthwhile to revisit the variability study discussed in Chapter 3 using TM rather than FM for all convolution computations. Note, however, that inverse problems using TM on the 24 data sets considered here required an average of 1241.1 minutes (20.7 hours) per data set. Therefore, running the inverse problems for all 1620 data sets considered in Chapter 3 would require about 1400 days of computer time.

Table 4.7: Absolute percent change in parameter estimates and percent change in costs for selected data sets when replacing Fenton method with trapezoid rule convolution method.

Data Set	Absolute % Change in:												% Change
	$\mathrm{E}[X_a]$	$\mathrm{SD}[X_a]$	c	$\mathrm{E}\left[T_0^{div}\right]$	$\mathrm{SD}\left[T_0^{div}\right]$	$\mathrm{E}\left[T^{div}\right]$	$\mathrm{SD}\left[T^{div}\right]$	$\mathrm{E}\left[T^{die}\right]$	$\mathrm{SD}\left[T^{die}\right]$	F_0	D_μ	D_σ	in Cost
1	38.40	24.06	5.48	0.90	1.59	3.22	297.04	86.40	3878.15	0.79	50.14	79.83	27.37
2	13.77	23.95	0.86	2.05	4.41	0.06	0.92	85.67	5359.90	5.02	3.63	2.49	9.70
3	38.66	24.04	5.51	1.05	1.83	3.25	294.38	88.74	3845.91	0.74	50.08	80.30	26.74
4	24.58	33.13	1.07	12.07	1328.19	61.40	97.93	30.14	21.12	56.00	3.68	20.24	-56.20
5	21.61	27.10	1.09	10.49	1300.12	60.68	97.94	37.90	35.06	50.53	3.83	18.96	-56.09
6	52.00	15.87	7.42	9.26	1229.55	60.42	92.36	36.64	48.60	36.31	50.49	135.41	-46.79
7	10.32	13.52	4.23	2.38	11.43	9.14	495.60	39.15	96.93	0.37	38.04	65.76	3.34
8	8.59	11.81	3.63	2.83	13.87	11.12	546.35	41.44	100.23	0.20	39.26	64.71	2.36
9	8.07	11.94	2.14	2.77	13.59	10.98	565.05	40.35	92.77	0.51	36.34	60.22	3.10
10	9.41	27.75	0.55	5.73	13.86	1.27	55.71	84.36	92.04	9.47	3.00	11.39	-29.40
11	9.96	31.47	0.45	5.19	22.01	2.99	86.00	78.93	457.92	10.74	18.08	21.82	-33.34
12	3.55	30.93	1.89	2.86	11.04	3.47	66.86	81.90	96.61	9.70	18.75	29.63	-34.43
13	11.05	27.45	0.73	0.40	1.07	0.40	4.98	4.36	23.73	0.39	1.32	10.41	-2.78
14	9.07	24.22	0.50	0.38	0.81	0.19	5.92	4.43	18.39	0.52	2.50	6.17	-2.38
15	7.58	24.96	0.25	0.31	0.48	0.01	1.10	6.46	5.08	0.65	4.81	0.35	-3.21
16	3.03	26.79	0.22	0.46	0.31	0.80	37.13	33.57	60.02	3.04	93.49	19.23	-14.56
17	0.33	23.10	1.14	0.82	1.45	1.64	39.54	31.20	64.23	1.99	82.70	22.90	-16.74
18	2.19	25.62	0.33	0.41	0.22	0.86	37.20	33.36	60.03	2.96	93.32	19.80	-14.17
19	7.36	10.05	2.29	3.01	2.08	9.89	6.07	36.98	96.64	2.17	13.48	23.45	3.95
20	9.43	17.17	0.98	0.03	2.31	2.72	0.54	20.60	27.28	0.79	4.59	2.84	-0.85
21	8.92	24.29	2.53	1.85	4.61	0.03	4.22	29.13	95.93	1.10	11.00	9.85	-3.91
22	31.58	63.91	6.18	2.00	6.51	2.55	7.24	49.46	96.96	4.21	16.91	12.11	-9.15
23	27.62	48.74	8.92	0.34	2.45	3.79	36.32	52.45	95.91	15.23	1.29	4.81	-16.53
24	17.66	33.72	10.48	0.27	0.22	2.35	37.73	51.03	96.68	13.93	6.24	2.94	-16.42
Median	9.69	24.62	1.51	1.92	3.43	2.86	38.63	38.52	94.34	2.56	15.20	19.52	-6.53

Chapter 5

Accounting for Asymmetric Division

We now relax the assumption of symmetric cell division and examine the more general case considered by Bocharov et al. [15] and Luzyanina et al. [32] in which each mother cell is assumed to distribute her cytoplasm in possibly uneven (but constant) proportions to two daughter cells during mitosis. That is, we assume there exists an $m \in (0, \frac{1}{2}]$ such that division of a mother cell results in two daughter cells that receive proportions m and $1 - m$, respectively, of the cytoplasm (and CFSE) that was contained in the mother cell. We begin by considering how this assumption of asymmetry in cell division, which has been described as "almost an axiom of cell biology" [44], affects the mathematical models presented in Chapter 2. We then study how incorporation of the asymmetric division parameter m into the specific mathematical model presented in Section 2.5 affects estimates of the model parameters.

5.1 Mathematical Model with Asymmetric Division

In order to obtain a PDE describing $n_i(t, x)$ in the case of possibly asymmetric cell divisions, we follow the same approach employed in Section 2.1; i.e., we consider the time rate of change of the total number of cells in generation i at time t with CFSE FI in the arbitrary interval $[x, x+\Delta x]$. Contributions (i) through (iv) to this rate of change, which are described in Section 2.1, are unaffected by asymmetry of cell divisions, but contribution (v) must be revisited. In this more general case, cells will enter the FI interval $[x, x + \Delta x]$ corresponding to generation i due to the division of cells in the previous generation $i - 1$ with FI in *two* distinct intervals. (Note that we must take Δx to be sufficiently small in order to ensure that the two FI intervals are, in fact, distinct.) To see this, note that any dividing cell in generation $i-1$ with FI in the interval $[\frac{1}{m}x, \frac{1}{m}(x + \Delta x)]$ will produce two cells, one of which will enter our control interval $[x, x + \Delta x]$ in generation i. This is because one of the daughter cells resulting from this division must have FI in the interval $[m \cdot \frac{1}{m}x, m \cdot \frac{1}{m}(x+\Delta x)] = [x, x+\Delta x]$. Similarly, any dividing cell in generation

$i-1$ with FI in the interval $[\frac{1}{1-m}x, \frac{1}{1-m}(x+\Delta x)]$ will produce two cells, one of which will enter our control interval $[x, x+\Delta x]$ in generation i.

Contribution (v), which describes the rate at which cells enter the control interval $[x, x+\Delta x]$ due to the division of cells in the previous generation, therefore consists of two terms. Each of these terms can be computed as the exponential birth rate times the number of cells in the relevant FI interval, so we have

$$\alpha_{i-1}(t) \cdot \int_{\frac{1}{m}x}^{\frac{1}{m}(x+\Delta x)} n_{i-1}(t,\xi)\, d\xi + \alpha_{i-1}(t) \cdot \int_{\frac{1}{1-m}x}^{\frac{1}{1-m}(x+\Delta x)} n_{i-1}(t,\xi)\, d\xi.$$

Applying the change of variables $\eta = m\xi$ in the first term and $\eta = (1-m)\xi$ in the second term, the expression becomes

$$\frac{1}{m} \cdot \alpha_{i-1}(t) \cdot \int_{x}^{x+\Delta x} n_{i-1}(t, \frac{1}{m}\eta)\, d\eta + \frac{1}{1-m} \cdot \alpha_{i-1}(t) \cdot \int_{x}^{x+\Delta x} n_{i-1}(t, \frac{1}{1-m}\eta)\, d\eta.$$

As was the case for symmetric cell division, it is important to note that contribution (v) *does not apply* in the case of cells in generation $i = 0$ because there is no previous generation from which cells can enter in that case.

Taking into account all the contributions, the time rate of change of the total number of cells in generation i at time t with FI in the region $[x, x+\Delta x]$ is

$$\begin{aligned}
\frac{d}{dt}\int_{x}^{x+\Delta x} n_i(t,\xi)\, d\xi &= \text{(i)} - \text{(ii)} - \text{(iii)} - \text{(iv)} + \text{(v)} \\
&= \Big[v(t)(x+\Delta x)n_i(t, x+\Delta x)\Big] - \Big[v(t)xn_i(t,x)\Big] \\
&\quad - \Big[\alpha_i(t) \cdot \int_{x}^{x+\Delta x} n_i(t,\xi)\, d\xi\Big] - \Big[\beta_i(t) \cdot \int_{x}^{x+\Delta x} n_i(t,\xi)\, d\xi\Big] \\
&\quad + \frac{1}{m}\alpha_{i-1}(t) \int_{x}^{x+\Delta x} n_{i-1}(t, \frac{1}{m}\eta)\, d\eta \\
&\quad + \frac{1}{1-m}\alpha_{i-1}(t) \int_{x}^{x+\Delta x} n_{i-1}(t, \frac{1}{1-m}\eta)\, d\eta.
\end{aligned}$$

Dividing by Δx on both sides of this equation and taking the limit as $\Delta x \to 0$ yields

$$\begin{aligned}
\frac{\partial}{\partial t}n_i(t,x) = v(t)\frac{\partial}{\partial x}\big[xn_i(t,x)\big] - \big(\alpha_i(t) + \beta_i(t)\big)n_i(t,x) + \frac{\alpha_{i-1}(t)}{m}n_{i-1}(t, \tfrac{1}{m}x) \\
+ \frac{\alpha_{i-1}(t)}{1-m}n_{i-1}(t, \tfrac{1}{1-m}x).
\end{aligned}$$

As mentioned above, the last two terms on the right hand side of this equation, which account

for recruitment of cells from the previous generation, must be omitted when $i = 0$. Thus, we obtain the model given in (2.1) with the recruitment terms $R_i(t,x)$ replaced by

$$R_i^{asym}(t,x) = \frac{\alpha_{i-1}(t)}{m} n_{i-1}(t, \tfrac{1}{m}x) + \frac{\alpha_{i-1}(t)}{1-m} n_{i-1}(t, \tfrac{1}{1-m}x) \tag{5.1}$$

for $i \geq 1$. Note that when $m = \frac{1}{2}$, $R_i^{asym}(t,x) = R_i(t,x)$ as defined in (2.2).

In keeping with the formulation discussed in Section 2.1, we propose that the solutions to the PDEs given in (2.1) with the modified recruitment terms given in (5.1) can still be factored as $n_i(t,x) = N_i(t)\bar{n}_i(t,x)$ (Equation 2.4), where $N_i(t)$ once again indicates the *number* of cells having completed i divisions at time t and $\bar{n}_i(t,x)$ describes the *distribution* of CFSE FI within that generation of cells at time t. The N_i's satisfy (2.5) and (2.6) as before, but each $\bar{n}_i$ now satisfies (2.7) and the *new* initial condition

$$\bar{n}_i(t_0,x) = \begin{cases} \dfrac{\Phi(x)}{N_0} & \text{for } i = 0, \\ \dfrac{1}{2m} \cdot \bar{n}_{i-1}(t_0, \tfrac{1}{m}x) + \dfrac{1}{2(1-m)} \cdot \bar{n}_{i-1}(t_0, \tfrac{1}{1-m}x) & \text{for } i \geq 1, \end{cases} \tag{5.2}$$

for all $x \geq 0$. Note that (5.2) reduces to (2.8) as expected when $m = \frac{1}{2}$. Note also that the recursive formula in (5.2) can be arrived at intuitively by recognizing that the two expressions of the form

$$\frac{1}{\lambda}\bar{n}_{i-1}(t_0, \tfrac{1}{\lambda}x)$$

(for $\lambda = m$ and $\lambda = 1-m$) are normalized pdfs representing densities of daughter cells in generation i (inheriting CFSE proportions $\lambda = m$ and $\lambda = 1-m$ from their predecessors) produced by dividing mother cells from generation $i-1$. The factor of $\frac{1}{2}$ in each of the terms in the recursive formula ensures that half of the daughter cells contain the proportions m and $1-m$, respectively, of the CFSE originally contained in the mother cells. We offer the following proposition concerning the factorability of solutions in the case of asymmetric cell divisions and provide a proof below.

Proposition 5.1. *Let $\{N_i(t)\}_{i=0}^{\infty}$ be a set of functions satisfying the system of weakly coupled ODEs given by* (2.5) *and the initial conditions given by* (2.6)*. Also, let $\{\bar{n}_i(t,x)\}_{i=0}^{\infty}$ be a set of functions such that each $\bar{n}_i$ satisfies the PDE* (2.7) *and the initial condition* (5.2) *for all $x \geq 0$. Then the solution to* (2.1) *with modified recruitment terms* (5.1) *and initial conditions* (2.3) *is given by* (2.4) *for $i \in \{0,1,2,\ldots\}$.*

Proof. The arguments for the case in which $i = 0$ are identical to those in the proof of Proposition 2.1, and therefore we do not repeat them here. For the cases in which $i \geq 1$, it is first necessary to obtain the solutions of (2.7) subject to the initial conditions (5.2). These solutions

can be obtained by the method of characteristics as shown in Section A.2 of the Appendix. From the expressions for these solutions it can be verified (see Lemma A.2) that

$$2\bar{n}_i(t,x) = \frac{1}{m}\bar{n}_{i-1}(t, \tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_{i-1}(t, \tfrac{1}{1-m}x) \tag{5.3}$$

for $i \geq 1$. We shall use this result in the following arguments.

For an index $i \geq 1$, taking the time derivative of (2.4) and then substituting (2.5) and (2.7) leads to

$$\frac{\partial}{\partial t}n_i(t,x) = -\Big(\alpha_i(t) + \beta_i(t)\Big)N_i(t)\bar{n}_i(t,x) + 2\alpha_{i-1}(t)N_{i-1}(t)\bar{n}_i(t,x) + v(t)\frac{\partial}{\partial x}\Big[xN_i(t)\bar{n}_i(t,x)\Big],$$

as was demonstrated in the proof for Proposition 2.1. Substituting (5.3) into the second term on the right hand side, we obtain

$$\begin{aligned}
\frac{\partial}{\partial t}n_i(t,x) &= -\Big(\alpha_i(t) + \beta_i(t)\Big)N_i(t)\bar{n}_i(t,x) \\
&\qquad + \alpha_{i-1}(t)N_{i-1}(t)\left[\frac{1}{m}\bar{n}_{i-1}(t, \tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_{i-1}(t, \tfrac{1}{1-m}x)\right] \\
&\qquad + v(t)\frac{\partial}{\partial x}\Big[xN_i(t)\bar{n}_i(t,x)\Big] \\
&= -\Big(\alpha_i(t) + \beta_i(t)\Big)n_i(t,x) + \frac{\alpha_{i-1}(t)}{m}n_{i-1}(t, \tfrac{1}{m}x) \\
&\qquad + \frac{\alpha_{i-1}(t)}{1-m}n_{i-1}(t, \tfrac{1}{1-m}x) + v(t)\frac{\partial}{\partial x}\Big[xn_i(t,x)\Big]
\end{aligned}$$

which is equivalent to (2.1) with modified recruitment terms (5.1) in the case $i \geq 1$. Substituting (2.6) and (5.2) into (2.4), the initial condition for an index $i \geq 1$ becomes

$$n_i(t_0,x) = N_i(t_0)\bar{n}_i(t_0,x) = \big[0\big] \cdot \left[\frac{1}{2m} \cdot \bar{n}_{i-1}(t_0, \tfrac{1}{m}x) + \frac{1}{2(1-m)} \cdot \bar{n}_{i-1}(t_0, \tfrac{1}{1-m}x)\right] = 0,$$

which is equivalent to (2.3) in the case $i \geq 1$.

Thus, we have verified that the solution to (2.1) with modified recruitment terms (5.1) and initial conditions (2.3) is given by (2.4) for $i \in \{0, 1, 2, \ldots\}$, provided the conditions stipulated in the proposition are met.

□

5.2 Asymmetric Division- and Label-Structured Cyton Model for Cell Densities

Now, as in Chapter 2, we incorporate the cyton model for cell numbers from Section 2.3 into our division- and label-structured model. In this case, we work with the *asymmetric* division- and label-structured model described in Section 5.1, the general form of which is given by (2.1) with modified recruitment terms (5.1). If we replace the sink and source terms in the right-hand sides of (2.1) with terms involving the cyton-based rates, we again obtain

$$\begin{aligned}
\frac{\partial n_0(t,x)}{\partial t} - v(t)\frac{\partial [x n_0(t,x)]}{\partial x} &= -\Big(n_0^{div}(t) + n_0^{die}(t)\Big)\bar{n}_0(t,x) && \text{(for } i = 0\text{)}, \\
\frac{\partial n_i(t,x)}{\partial t} - v(t)\frac{\partial [x n_i(t,x)]}{\partial x} &= \Big(2n_{i-1}^{div}(t) - n_i^{div}(t) - n_i^{die}(t)\Big)\bar{n}_i(t,x) && \text{for } i \geq 1.
\end{aligned}$$

(which is identical to Equation 2.14) as our division- and label-structured *cyton* model for cell densities in the case of possibly *asymmetric* cell division. We must be careful, however, to point out that each CFSE FI density $\bar{n}_i$ here satisfies a different initial condition than was posited in (2.8). Recall that the appropriate initial conditions in the case of asymmetric cell division are given by (5.2). As in the case of symmetric cell division, the solutions to the above system of partial differential equations are factorable. We offer this claim as the following proposition and provide a proof below.

Proposition 5.2. *Let $\{N_i(t)\}_{i=0}^{\infty}$ be a set of functions satisfying the cyton model* (2.11)*, where the initial condition N_0 is given by the relation shown in* (2.6)*. Also, let $\{\bar{n}_i(t,x)\}_{i=0}^{\infty}$ be a set of functions such that each $\bar{n}_i$ satisfies the PDE* (2.7) *and the initial condition* (5.2) *for all $x \geq 0$. Then the solution to* (2.14) *with initial conditions* (2.3) *is given by* (2.4) *for $i \in \{0, 1, 2, \ldots\}$.*

Proof. The arguments required here are identical to those given in the proof of Proposition 2.2, with the exception of arguments pertaining to initial conditions. Therefore, we address only issues relevant to initial conditions and refer the reader to the aforementioned proof for remaining details.

We begin by considering the case in which $i = 0$. Evaluating (2.4) at $t = t_0$ and substituting (2.11) and (5.2), the initial condition for the case $i = 0$ becomes

$$n_0(t_0, x) = N_0(t_0)\bar{n}_0(t_0, x) = \big[N_0\big] \cdot \left[\frac{\Phi(x)}{N_0}\right] = \Phi(x),$$

which is equivalent to (2.3) in the case $i = 0$.

Next, we consider the situation for $i \geq 1$. Evaluating (2.4) at $t = t_0$ and substituting (2.11)

and (5.2), the initial condition for the case $i \geq 1$ becomes

$$n_i(t_0, x) = N_i(t_0)\bar{n}_i(t_0, x) = \big[0\big] \cdot \left[\frac{1}{2m} \cdot \bar{n}_{i-1}(t_0, \tfrac{1}{m}x) + \frac{1}{2(1-m)} \cdot \bar{n}_{i-1}(t_0, \tfrac{1}{1-m}x)\right] = 0,$$

which is equivalent to (2.3) in the case $i \geq 1$.

Thus, we have verified that the solution to (2.14) and (2.3) is given by (2.4) for $i \in \{0, 1, 2, \ldots\}$, provided the conditions stipulated in the proposition are met. □

We remind the reader that (2.14) actually describes an entire class of models, as was pointed out in Chapter 2. To specify a particular model for further investigation, we must decide on forms for the distribution of the autofluorescence X_a, the (exponential) label decay rate $v(t)$, the cytons $\{(\phi_i(t), \psi_i(t))\}$, and the progressor fractions $\{F_i\}$.

5.3 Assumptions and Parameterization for a Specific Mathematical Model with Asymmetric Division

The specific asymmetric division cyton-based mathematical model we choose to examine satisfies exactly the same assumptions that were outlined for the twelve-parameter model described in Section 2.5. Thus, we have the 12 parameters of that model and the additional parameter $m \in (0, \frac{1}{2}]$, which indicates the degree of asymmetry associated with cell divisions as previously described. Table 5.1 shows the 13 parameters for our specific mathematical model. Precise details of our methods for computing numerical solutions for this model are provided in Appendix B. The interested reader should give special attention to Sections B.1.2 and B.4.2, which address numerical methods relevant to asymmetric division.

5.4 Importance of Accounting for Asymmetric Division

In this chapter, we are ultimately interested in determining whether or not refining our cell proliferation model to allow for asymmetric division is worthwhile. To make this determination, we perform parameter estimations using the same methodologies described in Section 2.6. Of course, in rereading that section, the structured densities $\tilde{n}(t, \tilde{x})$ and $\hat{n}(t, z)$ should be based on the thirteen-parameter asymmetric division model described in Section 5.3. Therefore, the vectors $\vec{q}_0$, $\hat{q}_{GLS}$, $\vec{q}_{typ}$, etc., referred to in Section 2.6 should be taken to be elements of $\mathbb{R}^{13}$ instead of $\mathbb{R}^{12}$. With this one modification, the generalized least squares parameter estimation scheme described in that section can be used here.

Parameter estimations for the asymmetric division model prove to be very computationally intensive, requiring an average of 851.3 minutes per inverse problem even when only considering

Table 5.1: Parameters for specific mathematical model with asymmetric division.

Number	Parameter	Description
1	$\mathrm{E}[X_a]$	mean autofluorescence
2	$\mathrm{SD}[X_a]$	std. dev. of autofluorescence
3	c	exponential decay rate for CFSE
4	m	asymmetric division constant
5	$\mathrm{E}\left[T_0^{div}\right]$	mean time to divide for cells in generation $i = 0$
6	$\mathrm{SD}\left[T_0^{div}\right]$	std. dev. of time to divide for cells in generation $i = 0$
7	$\mathrm{E}\left[T^{div}\right]$	mean time to divide for cells in later generations ($i \geq 1$)
8	$\mathrm{SD}\left[T^{div}\right]$	std. dev. of time to divide for cells in later generations ($i \geq 1$)
9	$\mathrm{E}\left[T^{die}\right]$	mean time to die for cells in later generations ($i \geq 1$)
10	$\mathrm{SD}\left[T^{die}\right]$	std. dev. of time to die for cells in later generations ($i \geq 1$)
11	F_0	progressor fraction for cells in generation $i = 0$
12	D_μ	mean of discrete normal distribution (used to compute F_i for $i \geq 1$)
13	D_σ	std. dev. of discrete normal distribution (used to compute F_i for $i \geq 1$)

cells that have divided $i_{\max} = 8$ or fewer times. (Timings provided here are based upon runs using MATLAB Release 2012a on a Dell Optiplex 990 with eight (8) 3.4 GHz Intel Core i7-2600 processors and 8 GB of 1333 MHz memory.) We discuss the choice of $i_{\max} = 8$ at length in Section 5.5. Because of the computational time required, we only apply our parameter estimation scheme to the the 24 data sets listed in Table 4.6. This leads to the results shown in Table 5.2. Since we are primarily interested in the asymmetric division parameter m, the table only shows estimates for that parameter. We see that in 8 of 24 cases (33.3% of the data sets), the parameter estimation scheme returns a value of *precisely* 0.5 for the parameter m. Recall that a value of $m = 0.5$ in the asymmetric division model yields model output equivalent to that of a *symmetric* division model. Moreover, in another 6 cases the value of m returned does not differ from 0.5 by more than 4 percent. That is, in $8 + 6 = 14$ cases (58.3% of the data sets), the parameter estimation scheme returns a value of m in the interval $[0.48, 0.5]$. This immediately suggests that asymmetry does *not* play an important role in the majority of the data sets considered.

To assess the degree to which the parameter m is influential in describing the 16 data sets for which m is not precisely 0.5, we perform model comparison tests similar to those described in Section 3.3.2. Here we once again consider two distinct mathematical models, both of which can be evaluated using the cost functional J given in (2.17). The first is the thirteen-parameter model described in Section 5.3 and the second is the nested model that results when the parameter m is fixed at the value 0.5 (which corresponds to perfectly symmetric cell divisions).

Table 5.2: Parameter estimates for m for 24 selected data sets. Data sets for which $m \neq 0.5$ are emphasized in boldface.

Data Set	Experiment	Data ID	m
1	Donor1/Vivid/CD4	10012	0.5000
2	Donor1/Vivid/CD4	10011	0.5000
3	Donor1/Vivid/CD4	11012	0.5000
4	Donor1/Vivid/CD4	21211	0.5000
5	**Donor1/Vivid/CD4**	**11211**	**0.4916**
6	Donor1/Vivid/CD4	11111	0.5000
7	Donor1/Vivid/CD8	21111	0.5000
8	Donor1/Vivid/CD8	20111	0.5000
9	Donor1/Vivid/CD8	10111	0.5000
10	**Donor1/Vivid/CD8**	**01200**	**0.4271**
11	**Donor1/Vivid/CD8**	**02000**	**0.4219**
12	**Donor1/Vivid/CD8**	**02200**	**0.4198**
13	**Donor2/Vivid/CD4**	**00222**	**0.4839**
14	**Donor2/Vivid/CD4**	**10222**	**0.4821**
15	**Donor2/Vivid/CD4**	**00212**	**0.4732**
16	**Donor2/Vivid/CD4**	**00100**	**0.4842**
17	**Donor2/Vivid/CD4**	**11110**	**0.4857**
18	**Donor2/Vivid/CD4**	**01100**	**0.4857**
19	**Donor2/Vivid/CD8**	**02001**	**0.4421**
20	**Donor2/Vivid/CD8**	**00002**	**0.4529**
21	**Donor2/Vivid/CD8**	**02020**	**0.4525**
22	**Donor2/Vivid/CD8**	**20211**	**0.4381**
23	**Donor2/Vivid/CD8**	**21211**	**0.4398**
24	**Donor2/Vivid/CD8**	**21111**	**0.4379**

To formulate our statistical hypotheses, we let $\mathcal{Q} \subset \mathbb{R}^{13}$ denote the set of all admissible parameters for the thirteen-parameter model and $\mathcal{Q}^H = \{\vec{q} \in \mathcal{Q} : H\vec{q} = \vec{c}\} \subset \mathcal{Q}$ be the set of admissible parameters for the nested model, where $H \in \mathbb{R}^{1\times 13}$ and $\vec{c} \in \mathbb{R}$. Using the parameter ordering suggested by Table 5.1 and setting

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and $\vec{c} = 0.5$ results in the nested model just described. We wish to test the null hypothesis that the "true" parameter vector $\vec{q}_0$ is in the restricted set $\mathcal{Q}^H$; i.e.,

$$H_0 : \vec{q}_0 \in \mathcal{Q}^H.$$

To proceed with the hypothesis test, we must define a test statistic. Following the approach

used in Section 3.3.2, let $\{N_k^j\}$ be a set of random variables as in (2.16) with corresponding realizations $\{n_k^j\}$ constituting observed data so that we can define the GLS estimators $\vec{q}_{GLS}$ and $\vec{q}_{GLS}^H$ as in (3.1) and GLS estimates $\hat{q}_{GLS}$ and $\hat{q}_{GLS}^H$ as in (3.2). As in Section 3.3.2, we note here that the inequality $J(\hat{q}_{GLS}^H; \{n_k^j\}) \geq J(\hat{q}_{GLS}; \{n_k^j\})$ should always hold because the estimate $\hat{q}_{GLS}^H$ is obtained by optimizing over a subset of $\mathcal{Q}$, while $\hat{q}_{GLS}$ is obtained by optimizing over all of $\mathcal{Q}$. Using the GLS estimators and estimates, we can define the test statistic $U(\{N_k^j\})$ with realization $\hat{U}(\{n_k^j\})$ as in (3.3) and (3.4), respectively. According to Banks and Tran [12], the test statistic U converges in distribution to a χ^2 distribution with $r = 1$ degree of freedom (where r is the number of constraints defined by the system $H\vec{q} = \vec{c}$) as $n \to \infty$. We use this statistic to test our null hypothesis.

As shown in Table 5.3, the costs associated with the optimal parameter vector $\hat{q}_{GLS}^H$ on the restricted set $\mathcal{Q}^H$ tend to be significantly greater than those associated with with the optimal parameter vector $\hat{q}_{GLS}$ on the set $\mathcal{Q}$. Note that the results shown in the table were obtained only for those data sets from Table 5.2 for which m was *not* equal to 0.5. We remark that for 3 of the selected data sets $J(\hat{q}_{GLS}^H; \{n_k^j\}) < J(\hat{q}_{GLS}; \{n_k^j\})$, which leads to a negative value for $\hat{U}(\{n_k^j\})$. This should not occur, and can probably be explained by imperfections in the optimization algorithms used for parameter estimation as discussed in Section 3.3.2. Based on the very low p-values resulting from the model comparison tests in the majority of cases, we should reject the null hypothesis and infer that the asymmetric division parameter m *is* important for describing the behavior of a population of proliferating T cells.

Thus, we have arrived at two seemingly contradictory results concerning the importance of incorporating asymmetric division into our model. In one third of the data sets we examined, our parameter estimation algorithm returned a value of $m = 0.5$, which indicates perfectly symmetric division. For the remaining data sets, however, our parameter estimation algorithm returned a value of m less than 0.5. Furthermore, if we fix m at 0.5 for these data sets, the GLS cost tends to be significantly greater. So, for two thirds of the data sets examined, allowing for asymmetric division does appear to lead to better agreement between the model and the data.

We hypothesize that the *apparent* importance of incorporating asymmetry that is implied by the model comparison experiments might be due to a type of experimental confounding. Recall that the statistical model we use to relate experimental data to our mathematical model is not a constant coefficient of variance model (cf. (2.16)). In fact, data points (histogram bins) with higher cell counts tend to contribute more to the cost functional given by (2.17), even despite the moderating effect of the weights. Also, the cell counts in the various bins tend to be larger in the later days of the experiment after the cells have had the opportunity to multiply their numbers through cell division. Therefore, our parameter estimation scheme tends to favor parameter sets that give good model fits (i.e., low residuals) for observations at later days (e.g., Days 4 and 5) over those that give good model fits for observations at earlier days (e.g.,

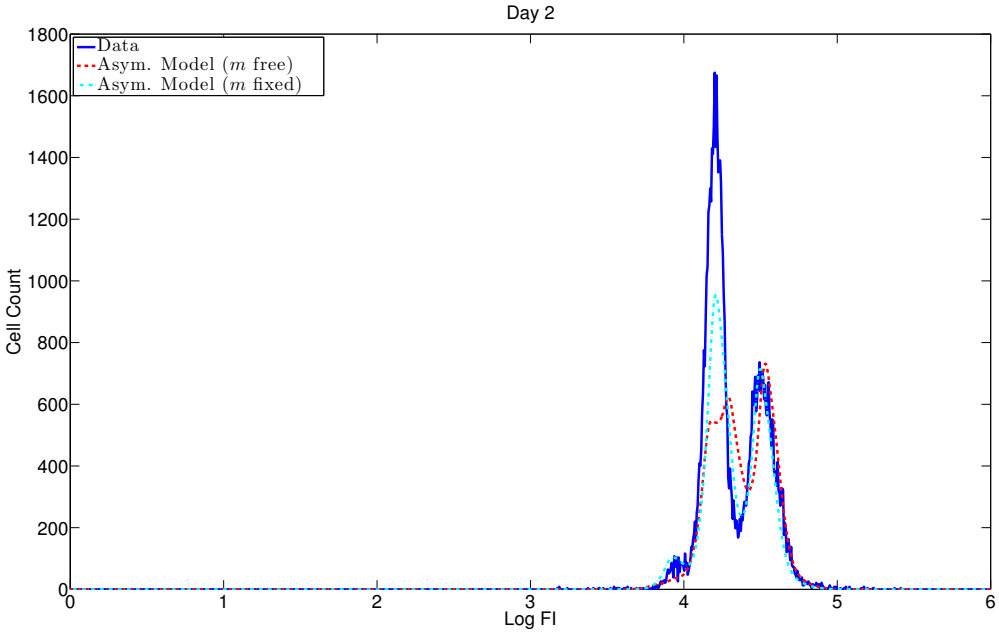
Table 5.3: Results of the model comparison test described in Section 5.4.

Data Set	$J(\hat{q}_{GLS}, \{n_k^j\})$	$J(\hat{q}_{GLS}^H, \{n_k^j\})$	$\hat{U}(\{n_k^j\})$	p-value
5	33235.4	13380.5	-2446.96	1
10	20460.4	20373	-17.4833	1
11	15997	20635.2	1187.6	0
12	17384.2	21583.3	989.366	0
13	9785.79	9893.49	45.0805	1.9×10^{-11}
14	9776.4	9932.86	65.5522	5.6×10^{-16}
15	11690.5	11423.5	-93.5471	1
16	13648.9	13749	30.0378	4.2×10^{-8}
17	13241	16056.1	870.828	0
18	14017.7	14104.8	25.4563	4.5×10^{-7}
19	9733.53	10234.3	210.737	0
20	10548.6	10566.6	6.97462	8.3×10^{-3}
21	9621.94	10349.5	309.72	0
22	10341.9	13651.4	1310.76	0
23	13075.5	13858.1	245.151	0
24	11323.2	14597.8	1184.55	0

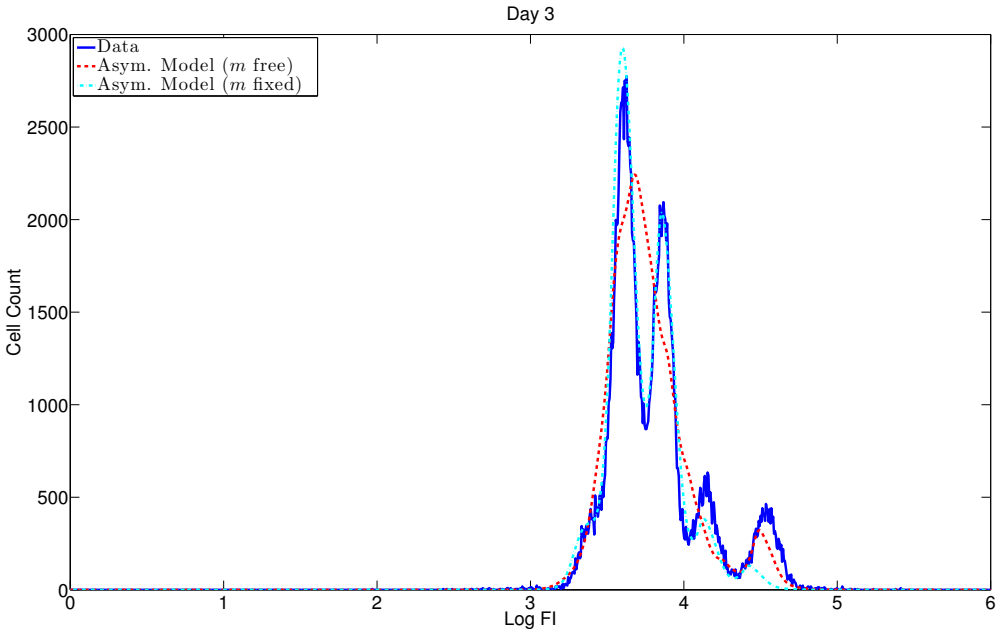
Days 2 and 3). As supporting evidence for this hypothesis, consider Figures 5.1 and 5.2. These figures show data from Data Set 12 of Table 5.2 along with the asymmetric model fits obtained (1) when m is free to take on any admissible value and (2) when m is fixed at 0.5. (Note that the latter model fit is equivalent to a symmetric division model fit.) In Figure 5.1, we clearly see that the symmetric division ("m fixed") model does a better job of capturing all the peaks in the data for the earlier days than does the asymmetric division ("m free") model. In examining Figure 5.2, it might even be argued that the symmetric division model captures the peaks in the data for the later days just as well as the asymmetric division model, but clearly the residuals produced by the asymmetric division model are smaller. Thus, despite the better overall "peak capturing" performance of the symmetric division model, the asymmetric division model is selected by the model comparison tests because it produces smaller residuals for the later days of observations. It is conceivable that the introduction of the additional parameter m allows one to obtain better fits using the asymmetric model simply because this additional degree of freedom allows one to obtain reasonably low residuals for the early days of observations while simultaneously obtaining very low residuals for the later days of observations. Perhaps applying an analysis of residuals (as described by Banks and Tran [12] and Banks et al. [3])

to the asymmetric division model would lead to a different and better statistical model upon which future parameter estimations could be based.

It should also be pointed out that, in some cases, neither the asymmetric nor the symmetric division model seem to do an adequate job of fitting CFSE FI data of the type described in Chapter 3. Figure 5.3, which shows data from Data Set 22 of Table 5.2 along with the asymmetric ("m free") and symmetric ("m fixed") model fits, illustrates this point. Notice that the fits obtained with both models are particularly poor at Days 2 and 3. While such poor fits could occur because of imperfections in the parameter estimation algorithm (e.g., optimization routines identifying local rather than global minima) or simply because some of our data sets contain unidentified experimental errors, it is also feasible that our models have yet to account for some important parameters governing T cell proliferation.

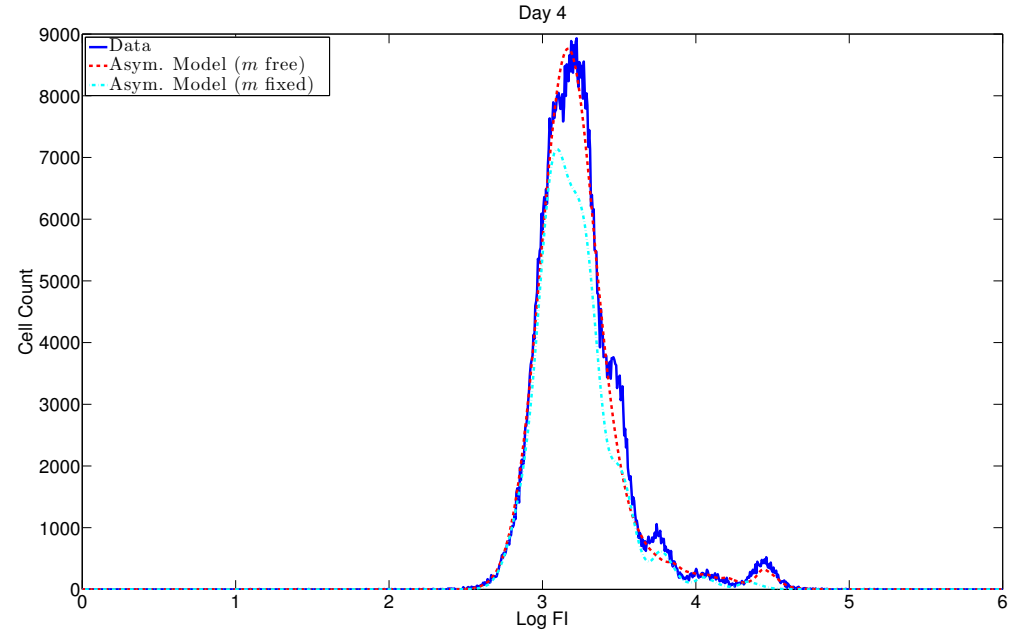


(a)

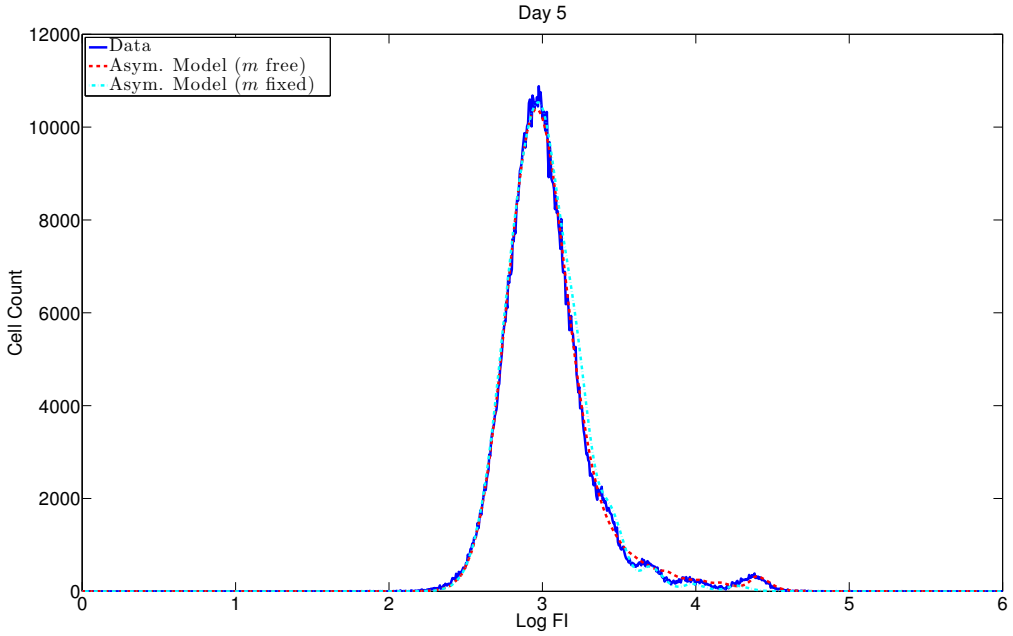


(b)

Figure 5.1: Plots of data from Data Set 12 of Table 5.2 and optimized model fits at (a) Day 2 and (b) Day 3.



(a)



(b)

Figure 5.2: Plots of data from Data Set 12 of Table 5.2 and optimized model fits at (a) Day 4 and (b) Day 5.

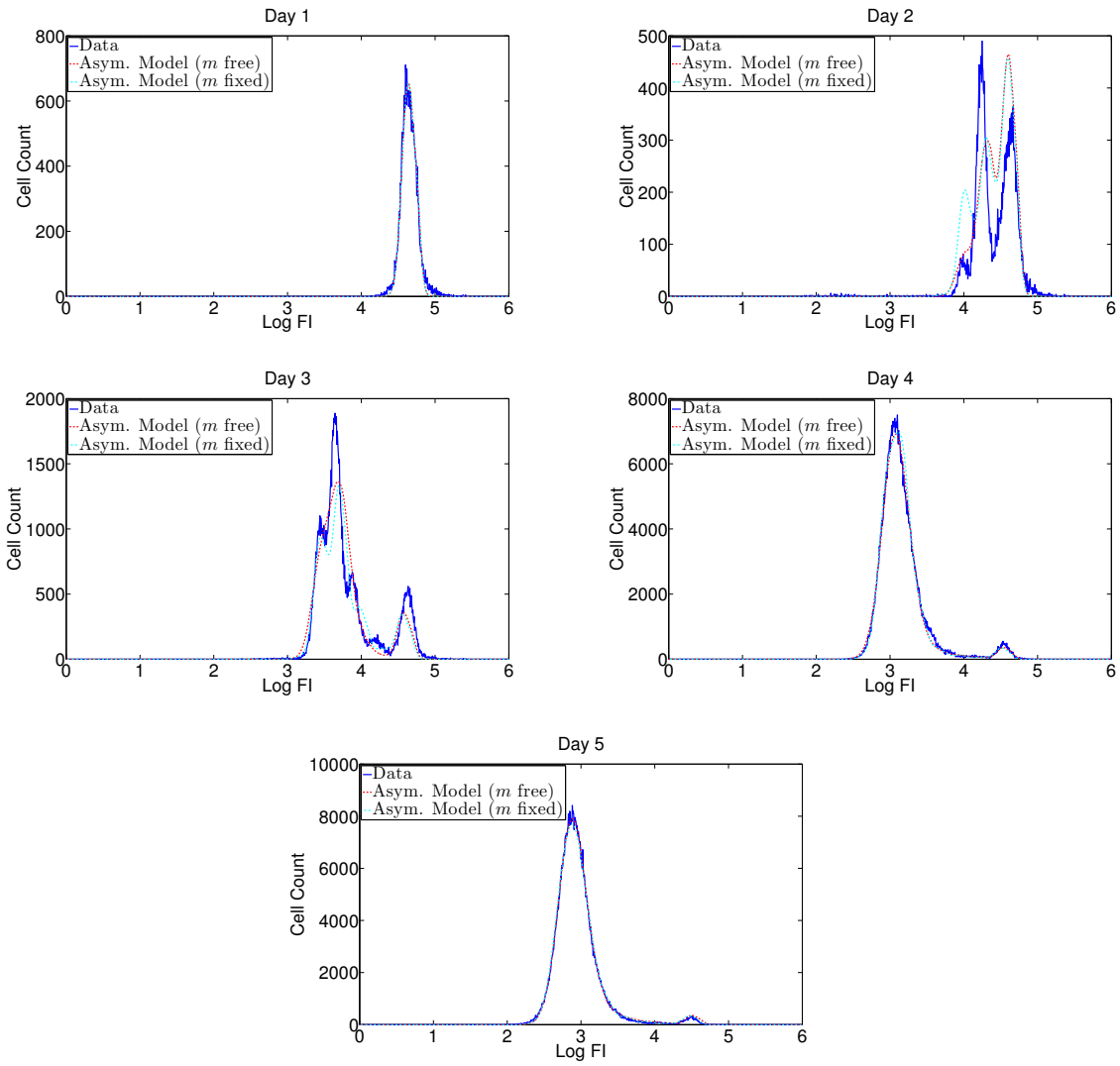


Figure 5.3: Plots of data from Data Set 22 of Table 5.2 and optimized model fits for Days 1 through 5.

5.5 On Choosing $i_{\max} = 8$ for the Asymmetric Division Model

Here we revisit a point to which we alluded briefly at the beginning of Section 5.4. For the asymmetric division model, we only perform computations for cells that have divided 8 or fewer times. That is, we only compute the solutions $\tilde{n}_i(t, \tilde{x})$ for $i \in \{0, \ldots, i_{\max}\}$, where $i_{\max} = 8$. We reduce the value for $i_{\max}$ from 16 (which was used for the symmetric division model of Chapter 2) because of the compounding expense of the recursive function calls that are required when computing numerical solutions for the asymmetric division model. (We provide details of

the relevant computational methods in Section B.1.2 of the Appendix.) This reduction, however, requires some justification.

Considering *most* of the data sets analyzed in Chapter 3 using our original symmetric division model, we find that it is very rare for the number of cells in generations $i = 9$ through $i = 16$ to make up more than 1% of the total number of cells in generations $i = 0$ through $i = 16$, even at Day 5 when the possibility of finding cells in these later generations is greatest. This is true except when considering data sets for Donor 2's CD8+ T cells, in which case the percentage of cells in generations 9 through 16 made up on average 9.5% of the total cell count at Day 5 when using ViViD dye and 7.6% when not using ViViD dye.

To further justify our choice of $i_{\max} = 8$, we conduct a test to determine how much parameter estimates change when when switching from $i_{\max} = 16$ to $i_{\max} = 8$. Note that we already have parameter estimates for 24 data sets based upon the asymmetric division model with $i_{\max} = 16$ and m fixed at 0.5 – these are the results for the trapezoid convolution method (TM) applied to the symmetric division model that were discussed in Section 4.3. We also already have parameter estimates for the same 24 data sets based upon the asymmetric division model with $i_{\max} = 8$ and m fixed at 0.5 – see Section 5.4. Thus, we can examine relative changes in the parameter estimates as $i_{\max}$ is adjusted from 16 to 8. In Table 5.4, we provide the results of our test. For this table, each percent change is calculated as

$$\frac{v_8 - v_{16}}{v_{16}} \times 100\%,$$

where v_8 and v_{16} denote the values obtained when using $i_{\max} = 8$ and $i_{\max} = 16$, respectively, and each *absolute* percent change is calculated as the absolute value of this quantity. We see that, with a few exceptions, the percent change in the cost is usually fairly small. Also, the median absolute percent change in each of the parameter estimates is less than 2%, except in the case of SD $\left[T^{die}\right]$ and D_σ. Note that rows of the table containing larger absolute percent changes in the parameter estimates (and the cost) correspond to data sets for which fixing m at 0.5 gave poor fits. Unfortunately, reattempting the parameter estimates using $i_{\max} = 16$ with m as a "free" parameter is not currently feasible due to the computational time required. Nevertheless, the results of this test tend to suggest that only small changes in most of the parameter estimates occur when switching from $i_{\max} = 16$ to $i_{\max} = 8$.

Table 5.4: Absolute percent change in parameter estimates and percent change in costs for selected data sets when switching from $i_{\max} = 16$ to $i_{\max} = 8$.

Data Set	Absolute % Change in:												% Change
	$\mathrm{E}[X_a]$	$\mathrm{SD}[X_a]$	c	$\mathrm{E}[T_0^{div}]$	$\mathrm{SD}[T_0^{div}]$	$\mathrm{E}[T^{div}]$	$\mathrm{SD}[T^{div}]$	$\mathrm{E}[T^{die}]$	$\mathrm{SD}[T^{die}]$	F_0	D_μ	D_σ	in Cost
1	0.09	0.50	0.04	0.01	0.01	0.20	1.92	0.47	1.08	0.01	0.21	0.48	-0.27
2	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.05	0.07	0.01	0.00	0.00	-0.02
3	0.10	0.53	0.05	0.02	0.01	0.20	1.96	0.44	1.01	0.01	0.24	0.49	-0.24
4	0.00	0.02	0.02	0.01	0.02	0.00	0.00	0.08	0.11	0.02	0.00	0.00	-0.10
5	0.39	4.86	0.16	2.15	4.04	0.08	1.08	47.67	98.30	5.33	0.46	0.78	-6.81
6	0.46	0.73	0.01	0.04	0.07	0.03	0.89	0.16	0.32	0.05	0.69	0.62	-0.11
7	1.52	0.00	0.59	0.24	0.66	1.17	3.33	1.26	0.93	0.06	0.89	5.15	-1.17
8	1.53	0.00	0.59	0.17	0.30	0.98	2.02	1.26	0.96	0.05	0.84	5.45	-1.25
9	1.39	0.00	0.30	0.18	0.36	1.01	2.18	1.45	1.17	0.02	0.62	5.59	-1.49
10	0.47	0.00	2.14	6.22	14.07	1.39	128.06	536.08	1251.61	9.37	4.41	39.57	35.25
11	2.14	0.20	1.40	5.90	21.80	2.69	638.99	687.55	1846.16	11.80	18.22	63.76	43.26
12	8.81	0.00	4.04	3.97	14.47	1.82	391.22	586.97	14688.75	10.41	18.92	82.89	45.51
13	1.37	2.87	0.04	0.09	0.24	0.04	0.01	1.37	10.20	0.21	0.08	10.78	-0.33
14	1.47	2.63	0.04	0.10	0.28	0.06	0.80	1.38	10.75	0.20	0.19	9.42	-0.39
15	1.56	3.71	0.76	0.27	0.26	1.61	75.73	85.38	368.76	0.07	54.72	39.47	12.12
16	0.63	0.79	0.03	0.03	0.02	0.00	0.07	0.27	2.33	0.21	0.89	2.68	-0.04
17	3.64	5.98	1.15	1.06	1.83	2.11	66.90	44.52	181.93	1.70	47.14	34.16	18.38
18	1.23	1.77	0.01	0.07	0.09	0.00	0.02	0.10	1.95	0.12	0.72	2.85	-0.00
19	9.80	14.09	0.21	0.01	3.31	2.60	0.20	0.93	0.84	0.99	0.27	16.05	-4.01
20	1.61	12.38	1.65	1.79	0.93	5.58	0.41	37.14	95.62	3.42	4.77	19.92	2.28
21	4.38	17.76	0.07	1.72	1.38	2.95	2.68	3.17	144.66	1.75	5.69	0.49	2.86
22	17.39	6.84	7.25	1.02	25.44	21.86	22.78	36.49	475.73	17.23	1.25	34.13	-3.46
23	14.40	2.33	4.23	3.27	20.19	23.29	10.62	34.16	321.64	6.80	19.89	43.51	3.72
24	17.30	8.12	3.65	2.13	24.12	24.15	6.75	57.22	27.03	15.45	18.32	29.05	8.48
median	1.50	1.28	0.25	0.21	0.51	1.09	1.99	1.41	10.47	0.21	0.87	7.51	-0.07

Chapter 6

Conclusions and Future Directions

In this final chapter, we summarize the results we've presented and discuss possible future investigations that might lead to improvements in the cell proliferation models and parameter estimation procedures we've described. Our most important discoveries were reported in Chapters 3, 4, and 5, so we discuss these results and some of their implications in turn in Section 6.1. This leads into a discussion of the opportunities for expanding on this body of work in Section 6.2.

6.1 Summary of Results

In Chapter 3, we offered evidence that 9 of the 12 parameters used to specify the symmetric division model of Section 2.5 can be estimated with relatively high reliability using data sets of the type described in Section 3.1. We then showed that our inability to reliably identify the remaining 3 parameters could be explained by the high degree of variability in observations made in the later days of the experiments. The increased variability seen in triplicate data sets after Day 3 was linked to the removal of some of the growth medium (and possibly some cells) and the subsequent addition of fresh growth medium that occurs starting at Day 3 according to the experimental protocol.

It is worth noting that we derived the mathematical models of Chapter 2 (and Chapter 5) according the principle of conservation of mass. Because cells are possibly removed from the culturing wells during the growth medium replenishment procedure that occurs starting at Day 3, the assumptions of our conservation laws may be violated. Furthermore, as pointed out in Section 3.4, the depletion and replenishment of nutrients available to the cells amounts to a non-constant environment. The cyton model we've incorporated into our structured PDE model tacitly assumes a constant environment for the cultures of proliferating cells, and we now clearly see that this is an unrealistic assumption. Thus, violations of conservation laws (for mass *and*

energy) could easily be responsible for the model misspecification pointed out in Section 2.6.

In Chapter 4, we described several numerical methods for computing the convolution integrals relevant to the cell proliferation models of Chapters 2 and 5. In particular, we examined how use of the Fenton method over the more precise (and more costly) trapezoid method may have affected the results discussed in Chapter 3. It was demonstrated that, while use of the Fenton method may have led to dramatically different estimates for many of the model parameters, estimates for 4 of the most important cell proliferation parameters were probably affected very little.

Chapter 5 contains a derivation of a new cell proliferation model that incorporates aspects of the cyton model of Hawkins et al. [28, 29], the separable PDE models of Schittler et al. [41], and the asymmetric division parameter m proposed by Bocharov et al. [15]. The model comparisons performed there did not provide conclusive evidence that including a provision for asymmetric division in our models consistently leads to improved fits for the data sets described in Section 3.1. In some cases, including the parameter m allowed for model fits with significantly lower costs (as evaluated using (2.17)), but in the majority of cases the estimated value of m was not substantially different from 0.5 (which corresponds to perfectly symmetric cell divisions). Upon visual inspection of some of the model fits obtained in Section 5.4, it was noticed that the selection of the asymmetric model over the symmetric model might be due to the form of the statistical model underlying the cost functional (2.17). It was then suggested that a reassessment of this statistical model might lead to different results.

6.2 Future Directions

In the previous section, we alluded to several opportunities for expanding on the work presented here. In particular, one might attempt derive new conservation law models similar to (2.1) or (2.14) that account for the conservation law violations identified in Section 6.1. This could possibly be accomplished by working with experimentalists to approximate the quantities of cells that are removed during growth medium replenishment. Alternatively, new experimental methods might be devised for culturing cells in such a way that the nutrients available in the growth medium are maintained at more constant levels. If new protocols for maintaining a more constant environment cannot be developed, then the cyton model incorporated into (2.14) may need to be modified or abandoned. In their current form, the cytons $\{(\phi_i(t), \psi_i(t))\}$ depend only on the time t since the last division occurred; i.e., they only depend on time in the frame of reference of an individual cell. In order to account for a non-constant environment, the cytons would need to also depend on time in the frame of reference of the experimenter. Adding this new dependence would likely complicate the cyton model presented in Section 2.3 to such an extent that the original benefits imparted (i.e., increased simplicity in the interpretation of the

model parameters and a reduced number of model parameters) would be eliminated. Thus, reverting to model (2.1) with its "real time" time-dependent division and death rates $\{\alpha_i(t)\}$ and $\{\beta_i(t)\}$ would probably be the best course in the event that new constant-environment experimental procedures cannot be suggested.

Considerable effort has already been devoted to the problem of parameterizing a model such as (2.1) that contains time-dependent rate functions $\{\alpha_i(t)\}$ and $\{\beta_i(t)\}$ for describing division and death. By conducting model comparison tests using the Akaike Information Criterion [17], Thompson found that representing each $\alpha_i(t)$ by a piecewise linear function and defining each $\beta_i(t)$ by a constant function led to the most parsimonious model when selecting from a panel of candidate model parameterizations [47]. Using Thompson's scheme, one must estimate the value of each $\alpha_i(t)$ at three carefully selected time points; therefore, when considering only generations $i = 0$ through $i = 6$, $7 \times 3 = 21$ parameters are required to represent the division rates and 7 more are required for the death rates. Although Thompson was able to obtain good fits to summary histogram FI data with this model formulation, it is unlikely that that all the parameters involved can be reliably estimated.

For the data sets described in Chapter 3, recall that observations were made at 5 discrete time points on roughly 24-hour intervals over a 5-day period. In order to reliably estimate all the parameters involved when discretely parameterizing $\{\alpha_i(t)\}$ and $\{\beta_i(t)\}$ as described above, it seems likely that observations would need to be made with greater frequency. In discussing this possibility with our experimental collaborators, we learned that increasing the number of observations may not be practical for a variety of reasons. One of the most limiting factors is the volume of blood that can be safely drawn from a donor. A relatively large quantity of blood must be drawn in order to provide a sufficient number of PBMCs to conduct even the 5-day time series experiments of Chapter 3, and this study was conducted using blood from *healthy* donors. In studies conducted for diseased donors, the quantity of blood that can be safely removed may be even less.

It is worth noting that the triplicate data collected for the variability study of Chapter 3 could be used in an entirely different way. In particular, one could attempt to obtain parameter estimates using *all* 15 data points for a given experimental condition. That is, instead of choosing one of the 243 possible 5-day time series data sets, one could use all 3 measurements from all 5 days together to find a parameter vector that gives the best overall fit in the least-squares sense. As a similar alternative, one could apply k-fold cross validation [24]. In this approach, one would systematically exclude (one at a time) each of the 12 data points from the later days (i.e., Days 2 through 5) of the experiment and would use the 11 remaining data points plus the 3 data points from Day 1 (for the initial condition) to obtain a parameter estimate. Then one could analyze the ability of the model to "predict" the information contained in the excluded data point.

As another possibility for moving forward with this work, we alluded to statistical model revisions in the previous section. A number of approaches to accommodating model discrepancy, from the relatively simple (e.g., polynomial model discrepancy terms) to the relatively complicated (e.g., Gaussian process representations) are addressed by Smith [45] and could be applied here. Given the results of Section 3.2, it is now clear that our experimental process leads to a greater potential for measurement error in the later days (especially after Day 3) and it appears that this error is introduced through the growth medium replenishment process. Note that the statistical model given in Section 2.6 can be expressed as

$$N_k^j(\vec{q}_0) = I[\hat{n}](t_j, z_k; \vec{q}_0) + \epsilon_k^j \sqrt{\frac{B}{b_j} I[\hat{n}](t_j, z_k; \vec{q}_0)},$$

where each ϵ_k^j is standard normal random variable (i.e., $\epsilon_k^j \sim \mathcal{N}(0,1)$). This statistical model could be easily modified by selecting the variance of each ϵ_k^j in such a way that the total relative variation in N_k^j would be larger in the later days of the experiment (i.e., for $j \in \{3,4,5\}$). One could use the relative variation measures we describe in Section 3.2 as a guide for choosing these variances.

In all the models we have considered thus far, note that we have assumed that cells of a specific type (e.g., CD4+ and CD8+ T cells) operate independently of one another. In fact, it is well known that the proliferation of CD8+ T cells depends in many cases on interactions with CD4+ T cells [14, 19]. Furthermore, other intermediary cells (antigen-presenting cells such as dendritic cells) often facilitate the communication between CD4+ and CD8+ T cells [14]. We suggest that investigations of a two-population (or possibly a three-population) model might lead to improvements of model fits. In such a model, the behavior of CD4+ and CD8+ T cells would be considered simultaneously, and the birth rate terms for CD8+ T cells would include a dependence on both CD8+ *and* CD4+ numbers.

Ultimately, we would like to be able to apply our models and methods to more interesting and practical scenarios, such as investigations of Gag-stimulated cell cultures from HIV-positive donors. For the data we considered in Chapters 3 and 5, recall that cells were stimulated to divide using the mitogen PHA. Because PHA is a nonspecific T cell mitogen, the response rate of T cells stimulated to divide with PHA is *much higher* than that of T cells stimulated with a group-specific antigen. In fact, in examining some preliminary data provided by our collaborators at Universitat Pompeu Fabra, we found that the response rate of T cells collected from HIV-positive donors and stimulated with the Gag protein can be 1% or less. Compare this with the 50 to 90% response rate observed for PHA-stimulated T cells.

Thus, a variety of opportunities and challenges remain to be considered in the area of cell proliferation modeling. The models and experimental procedures described here have allowed

us to glimpse a number of quantifiable features of populations of proliferating cells, and as more information becomes available from the biological research community we believe that the interpretive framework we've described will provide a firm foundation for further investigations.

REFERENCES

[1] H.T. Banks, F. Charles, M. Doumic, K.L. Sutton, and W.C. Thompson, Label structured cell proliferation models, CRSC-TR10-10, North Carolina State University, June 2010; *Appl. Math. Letters* **23** (2010), 1412–1415.

[2] H.T. Banks, A. Choi, T. Huffman, J. Nardini, L. Poag, W.C. Thompson, Quantifying CFSE label decay in flow cytometry data, CRSC-TR12-20, North Carolina State University, December 2012; *Applied Math. Letters* **26** (2013), 571–577.

[3] H.T. Banks, S. Hu, W.C. Thompson, *Modeling and Inverse Problems in the Presence of Uncertainty*, CRC Press, Boca Raton, 2014.

[4] H.T. Banks, D.F. Kapraun, K.G. Link, W.C. Thompson, Cristina Peligero, Jordi Argilaguet, and Andreas Meyerhans, Experimental and biological variability in CFSE-based flow cytometry data, CRSC-TR13-10, North Carolina State University, September 2013.

[5] H.T. Banks, D.F. Kapraun, W.C. Thompson, C. Peligero, J. Argilaguet, and A. Meyerhans, A novel statistical analysis and interpretation of flow cytometry data, CRSC-TR12-23, North Carolina State University, March 2013; *J. Biological Dynamics* **7** (2013), 96–132.

[6] H.T. Banks, D.F. Kapraun, K.G. Link, W.C. Thompson, C. Peligero, J. Argilaguet, and A. Meyerhans, Analysis of variability in estimates of cell proliferation parameters for cyton-based models using CFSE-based flow cytometry data, CRSC-TR13-14, North Carolina State University, November 2013; *J. Inverse and Ill-posed Problems*, accepted.

[7] H.T. Banks, Z.R. Kenz, and W.C. Thompson, An Extension of RSS-based Model Comparison Tests for Weighted Least Squares, CRSC-TR12-18, North Carolina State University, August 2012; *Intl. J. Pure and Appl. Math.* **79** (2012).

[8] H.T. Banks, K.L. Sutton, W.C. Thompson, G. Bocharov, M. Doumic, T. Schenkel, J. Argilaguet, S. Giest, C. Peligero, and A. Meyerhans, A new model for the estimation of cell proliferation dynamics using CFSE data, CRSC-TR11-05, North Carolina State University, Revised July 2011; *J. Immunological Methods* **373** (2011), 143–160; DOI:10.1016/j.jim.2011.08.014.

[9] H.T. Banks, K.L. Sutton, W.C. Thompson, G. Bocharov, D. Roose, T. Schenkel, and A. Meyerhans, Estimation of cell proliferation dynamics using CFSE data, CRSC-TR09-17, North Carolina State University, August 2009; *Bull. Math. Biol.* **70** (2011), 116–150.

[10] H.T. Banks and W.C. Thompson, A division-dependent compartmental model with cyton and intracellular label dynamics, CRSC-TR12-12, North Carolina State University, May 2012; *Intl. J. Pure and Appl. Math.* **77** (2012), 119–147.

[11] H.T. Banks and W.C. Thompson, Mathematical models of dividing cell populations: Application to CFSE data, CRSC-TR12-10, North Carolina State University, April 2012; *J. Math. Modeling of Natural Phenomena* **7** (2012), 24–52.

[12] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton, 2009.

[13] N.C. Beaulieu, A.A. Abu-Dayya, and P.J. McLane, Estimating the distribution of a sum of independent lognormal random variables, *IEEE Trans. Commun.* **43** (1995), 2869–2873.

[14] M.J. Bevan, Helping the CD8+ T-cell response, *Nat. Rev. Immunol.* **4** (2004), 595–602.

[15] G. Bocharov, T. Luzyanina, J. Cupovic, and B. Ludewig, Asymmetry of cell division in CFSE-based lymphocyte proliferation analysis, *Front. Immunol.* **4** (2013), published online.

[16] A.W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Oxford UP, New York, 1997.

[17] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd Edition), Springer, New York, 2002.

[18] G. Casella and R.L. Berger, *Statistical Inference* (2nd Edition), Duxbury, 2002.

[19] F. Castellino and R.N. Germain, Cooperation between CD4+ and CD8+ T cells: When, where, and how, *Nat. Rev. Immunol.* **24** (2006), 519–540.

[20] M. Davidian and D.M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London, 2000.

[21] R.D. De Veaux, P.F. Velleman, D.E. Bock, *Stats: Models and Data* (3rd Edition), Pearson, Boston, 2012.

[22] R.J. De Boer and A.S. Perelson. Estimating division and death rates from CFSE data, *J. Comp. Appl. Math.* **184** (2005), 140-164.

[23] E.K. Deenick, A.V. Gett, and P.D. Hodgkin, Stochastic model of T cell proliferation: A calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival, *J. Immunology* **170** (2003), 4963–4972.

[24] G. Dougherty, *Pattern Recognition and Classification: An Introduction*, Springer, New York, 2013.

[25] L.F. Fenton, The sum of log-normal probability distributions in scatter transmission systems, *Trans. Commun. Syst.* **8** (1960), 57–67.

[26] A.V. Gett and P.D. Hodgkin, A cellular calculus for signal integration by T cells, *Nature Immunology* **1** (2000), 239–244.

[27] J. Hasenauer, D. Schittler, and F. Allgöwer, Analysis and simulation of division- and label-structured population models: a new tool to analyze proliferation assays, *Bull. Math. Biol.* **74** (2012), 2692–2732.

[28] E.D. Hawkins, M. Hommel, M.L. Turner, F. Battye, J. Markham and P.D. Hodgkin, Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data, *Nature Protocols* **2** (2007), 2057–2067.

[29] E.D. Hawkins, M.L. Turner, M.R. Dowling, C. van Gend, and P.D. Hodgkin, A model of immune regulation as a consequence of randomized lymphocyte division and death times, *Proc. Natl. Acad. Sci.* **104** (2007), 5032–5037.

[30] Chia-Lu Ho, Calculating the mean and variance of power sums with two log-normal components, *IEEE Trans. Veh. Technol.* **44** (1995), 756–762.

[31] E.L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1975.

[32] T. Luzyanina, J. Cupovic, B. Ludewig, and G. Bocharov, Mathematical models for CFSE labelled lymphocyte dynamics: asymmetry and time-lag in division, *J. Math. Biol.*, 13 December 2013 (published online ahead of print).

[33] T. Luzyanina, D. Roose, and G. Bocharov, Distributed parameter identification for a label-structured cell population dynamics model using CFSE histogram time-series data, *J. Math. Biol.* **59** (2009), 581–603.

[34] T. Luzyanina, D. Roose, T. Schenkel, M. Sester, S. Ehl, A. Meyerhans, and G. Bocharov, Numerical modelling of label-structured cell population growth using CFSE distribution data, *Theoretical Biology and Medical Modelling* **4** (2007), published online.

[35] A.B. Lyons, J. Hasbold, and P.D. Hodgkin, Flow cytometric analysis of cell division history using dilution of carboxyfluorescein diacetate succinimidyl ester, a stably integrated fluorescent probe, *Methods in Cell Biology* **63** (2001), 375–398.

[36] A.B. Lyons and C.R. Parish, Determination of lymphocyte division by flow cytometry, *J. Immunological Methods* **171** (1994), 131–137.

[37] G. Matera, M. Lupi, and P. Ubezio, Heterogeneous cell response to topotecan in a CFSE-based proliferation test, *Cytometry A* **62** (2004), 118–128.

[38] R.E. Nordon, M. Nakamura, C. Ramirez, and R. Odell, Analysis of growth kinetics by division tracking, *Immunology and Cell Biology* **77** (1999), 523–529.

[39] B. Quah, H. Warren, and C. Parish, Monitoring lymphocyte proliferation in vitro and in vivo with the intracellular fluorescent dye carboxyfluorescein diacetate succinimidyl ester, *Nature Protocols* **2** (2007), 2049–2056.

[40] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*, 2nd ed., Springer, New York, 2007.

[41] D. Schittler, J. Hasenauer, and F. Allgöwer, A generalized model for cell proliferation: Integrating division numbers and label dynamics, *Proc. Eighth International Workshop on Computational Systems Biology (WCSB 2011)*, June 2011, Zurich, Switzerland, p. 165–168.

[42] S.C. Schwartz and Y.S. Yeh, On the distribution function and moments of power sums with log-normal components, *Bell System Technical Journal* **61** (1982), 1441–1462.

[43] G.A. Seber and C.J. Wild, *Nonlinear Regression*, Wiley, Hoboken, NJ, 2003.

[44] R. Sennerstam, Partition of protein (mass) to sister cell pairs at mitosis: a re-evaluation, *J. Cell. Sci.* **90** (1988), 301–306.

[45] R.C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, SIAM, Philadelphia, 2014.

[46] M. Takaki, MATLAB function `SYSumLogNormal`, `http://www.snowelm.com/~t/doc/tips/20110902.en.html`, 2014.

[47] W.C. Thompson, *Partial Differential Equation Modeling of Flow Cytometry Data from CFSE-based Proliferation Assays,* Ph.D. Dissertation, Dept. of Mathematics, North Carolina State University, Raleigh, December 2011.

[48] S.A. Weston and C.R. Parish, New fluorescent dyes for lymphocyte migration studies: Analysis by flow cytometry and fluorescence microscopy, *J. Immunological Methods* **133** (1990), 87-97.

APPENDICES

Appendix A

Supporting Mathematical Arguments and Proofs

A.1 Solution of (2.7) and (2.8) by the Method of Characteristics

To simplify notation in the work that follows, we let $\rho(t,x) \equiv \bar{n}_i(t,x)$. Then, (2.7) can be written as

$$\frac{\partial}{\partial t}\rho(t,x) - v(t)\frac{\partial}{\partial x}\Big[x\rho(t,x)\Big] = 0.$$

Expanding the derivative with respect to x yields

$$\frac{\partial}{\partial t}\rho(t,x) - v(t)\Big[x\frac{\partial}{\partial x}\rho(t,x) + \rho(t,x)\Big] = 0,$$

which can be expressed as

$$\rho_t - v(t)x\rho_x - v(t)\rho = 0 \tag{A.1}$$

using the subscript notation for partial derivatives. We remark that, for our purposes, the domain for this PDE is $\{(t,x) : t \geq t_0, x \geq 0\}$, as is made clear in Section 2.1.

The goal of applying the method of characteristics to the first-order linear PDE (A.1) is to change from the coordinates (t,x) to a new coordinate system (s,x_0) so that the PDE becomes an ODE along the characteristic curves $\big\{\big(x(s),t(s)\big) : 0 < s < \infty\big\}$ in the x-t plane. The variable x_0 represents the initial value of x for each of the various characteristic curves, and it coincides with the point at which a given characteristic curve intersects the line $t = t_0$ in the x-t plane. If we let $\frac{dt}{ds}$ and $\frac{dx}{ds}$ equal the coefficients of ρ_t and ρ_x, respectively, in (A.1), we have the two initial value problems

$$\frac{dt}{ds} = 1, \qquad t(0) = t_0, \tag{A.2}$$

and

$$\frac{dx}{ds} = -v(t)x, \qquad x(0) = x_0. \tag{A.3}$$

The solution to (A.2) is

$$t(s) = t_0 + s,$$

and the solution to (A.3) can be obtained as follows.

$$\begin{aligned}
& \frac{dx}{ds} &=& -v(t)x \\
\Longrightarrow \quad & \frac{dx}{x} &=& -v(t)\,ds \\
\Longrightarrow \quad & \frac{dx}{x} &=& -v(t)\,dt \\
\Longrightarrow \quad & \int_{x_0}^{x^*} \frac{dx}{x} &=& -\int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \log x\Big|_{x_0}^{x^*} &=& -\int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \log x^* &=& \log x_0 - \int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & x^* &=& x_0 \cdot \exp\left[-\int_{t_0}^{t^*} v(t)\,dt\right] \\
\Longrightarrow \quad & x(s) &=& x_0 \cdot \exp\left[-\int_{t_0}^{t(s)} v(u)\,du\right].
\end{aligned}$$

This last equation implies the relation

$$x_0(s) = x(s) \cdot \exp\left[\int_{t_0}^{t(s)} v(u)\,du\right]. \tag{A.4}$$

Now, using (A.1), (A.2), and (A.3), we can write

$$\rho_t \frac{dt}{ds} + \rho_x \frac{dx}{ds} = v(t)\rho,$$

which implies

$$\frac{d\rho}{ds} = v(t)\rho, \tag{A.5}$$

where $\rho = \rho(s) = \rho(t(s), x(s))$ and $t = t(s)$. Also, the initial condition (2.8) implies

$$\rho_0 = \rho(0) = \frac{2^i \Phi(2^i x_0)}{N_0}. \tag{A.6}$$

Then (A.5) and (A.6) give us yet another initial value problem, which can be solved as follows.

$$\begin{aligned}
& \frac{d\rho}{ds} &=& v(t)\rho \\
\Longrightarrow \quad & \frac{d\rho}{\rho} &=& v(t)\,ds \\
\Longrightarrow \quad & \frac{d\rho}{\rho} &=& v(t)\,dt \\
\Longrightarrow \quad & \int_{\rho_0}^{\rho^*} \frac{d\rho}{\rho} &=& \int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \log\rho\Big|_{\rho_0}^{\rho^*} &=& \int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \log\rho^* &=& \log\rho_0 + \int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \rho^* &=& \rho_0 \cdot \exp\left[\int_{t_0}^{t^*} v(t)\,dt\right] \\
\Longrightarrow \quad & \rho(s) &=& \rho_0 \cdot \exp\left[\int_{t_0}^{t(s)} v(u)\,du\right] \\
\Longrightarrow \quad & \rho(s) &=& \frac{2^i\Phi(2^i x_0(s))}{N_0} \cdot \exp\left[\int_{t_0}^{t(s)} v(u)\,du\right].
\end{aligned}$$

Substituting (A.4) into the last equation above, we obtain

$$\rho(t,x) = \rho(t(s),x(s)) = \rho(s) = \frac{2^i}{N_0}\Phi\left(2^i x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right].$$

Recalling that $\rho(t,x) \equiv \bar{n}_i(t,x)$, we see that

$$\bar{n}_i(t,x) = \frac{2^i}{N_0}\Phi\left(2^i x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right] \tag{A.7}$$

is the solution of (2.7) and (2.8) for each $i \in \{0,1,2,\ldots\}$.

Lemma A.1. *For the solutions* (A.7) *of* (2.7) *and* (2.8)*, the relation*

$$\bar{n}_i(t,x) = 2\bar{n}_{i-1}(t,2x)$$

holds for all $i \in \{1,2,\ldots\}$.

Proof. Using (A.7), it is easy to see that

$$\begin{aligned} 2\bar{n}_{i-1}(t,2x) &= 2 \cdot \frac{2^{i-1}}{N_0}\Phi\left(2^{i-1}\cdot 2x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right)\cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right] \\ &= \frac{2^i}{N_0}\Phi\left(2^i x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right)\cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right] \\ &= \bar{n}_i(t,x) \end{aligned}$$

for all $i \in \{1, 2, \ldots\}$.

□

A.2 Solution of (2.7) and (5.2) by the Method of Characteristics

Note that here we solve the same first-order linear PDE which we solved in Appendix A.1, but we now consider a different initial condition. As before, we let $\rho(t,x) \equiv \bar{n}_i(t,x)$ to simplify notation, and in fact we proceed with identical arguments up through the point at which we obtained (A.5) in Appendix A.1.

For the case $i = 0$, note that (5.2) and (2.8) are equivalent expressions, so the solution for $\bar{n}_0(t,x) = \rho(t,x)$ will be the same as that obtained in Appendix A.1. For $i \geq 1$, however, the solutions will be different. We consider that case presently.

For $i \geq 1$, the initial condition (5.2) implies

$$\rho_0 = \rho(0) = \frac{1}{2}\cdot\left[\frac{1}{m}\bar{n}_{i-1}(t_0, \tfrac{1}{m}x_0) + \frac{1}{(1-m)}\bar{n}_{i-1}(t_0, \tfrac{1}{1-m}x_0)\right]. \tag{A.8}$$

Then (A.5) and (A.8) give us an initial value problem that can be solved as follows.

$$\begin{aligned}
& \frac{d\rho}{ds} &=& v(t)\rho \\
\Longrightarrow \quad & \frac{d\rho}{\rho} &=& v(t)\,ds \\
\Longrightarrow \quad & \frac{d\rho}{\rho} &=& v(t)\,dt \\
\Longrightarrow \quad & \int_{\rho_0}^{\rho^*} \frac{d\rho}{\rho} &=& \int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \log\rho\Big|_{\rho_0}^{\rho^*} &=& \int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \log\rho^* &=& \log\rho_0 + \int_{t_0}^{t^*} v(t)\,dt \\
\Longrightarrow \quad & \rho^* &=& \rho_0 \cdot \exp\left[\int_{t_0}^{t^*} v(t)\,dt\right] \\
\Longrightarrow \quad & \rho(s) &=& \rho_0 \cdot \exp\left[\int_{t_0}^{t(s)} v(u)\,du\right] \\
\Longrightarrow \quad & \rho(s) &=& \frac{1}{2}\left[\frac{1}{m}\bar{n}_{i-1}(t_0, \tfrac{1}{m}x_0) + \frac{1}{1-m}\bar{n}_{i-1}(t_0, \tfrac{1}{1-m}x_0)\right] \cdot \exp\left[\int_{t_0}^{t(s)} v(u)\,du\right].
\end{aligned}$$

Substituting (A.4) into the last equation above, we obtain

$$\begin{aligned}
\rho(t,x) &= \rho(t(s),x(s)) = \rho(s) \\
&= \frac{1}{2}\left[\frac{1}{m}\bar{n}_{i-1}\left(t_0, \frac{1}{m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right)\right. \\
&\qquad \left. + \frac{1}{1-m}\bar{n}_{i-1}\left(t_0, \frac{1}{1-m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right)\right] \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right].
\end{aligned}$$

Recalling that $\rho(t,x) \equiv \bar{n}_i(t,x)$, we see that the solutions $\bar{n}_i(t,x)$ of (2.7) and (5.2) obey the recursive relation

$$\begin{aligned}
\bar{n}_i(t,x) = \frac{1}{2}&\left[\frac{1}{m}\bar{n}_{i-1}\left(t_0, \frac{1}{m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right)\right. \\
&\left. + \frac{1}{1-m}\bar{n}_{i-1}\left(t_0, \frac{1}{1-m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right)\right] \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right] \qquad \text{(A.9)}
\end{aligned}$$

for $i \geq 1$.

Lemma A.2. *For the solutions* (A.9) *of* (2.7) *and* (5.2)*, the relation*

$$2\bar{n}_i(t,x) = \frac{1}{m}\bar{n}_{i-1}(t, \tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_{i-1}(t, \tfrac{1}{1-m}x)$$

holds for all $i \in \{1, 2, \ldots\}$.

Proof. We will prove this claim using mathematical induction. First, we show that the claim holds for $i = 1$. Note that (A.7), which holds for $i = 0$, implies

$$\bar{n}_0(t,x) = \frac{1}{N_0}\Phi\left(x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right], \tag{A.10}$$

so

$$\begin{aligned}
\bar{n}_0(t_0,\xi) &= \frac{1}{N_0}\Phi\left(\xi \cdot \exp\left[\int_{t_0}^{t_0} v(u)\,du\right]\right) \cdot \exp\left[\int_{t_0}^{t_0} v(u)\,du\right] \\
&= \frac{1}{N_0}\Phi\left(\xi \cdot \exp\left[0\right]\right) \cdot \exp\left[0\right] \\
&= \frac{1}{N_0}\Phi(\xi)
\end{aligned}$$

and thus

$$\bar{n}_0\left(t_0, \frac{1}{m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) = \frac{1}{N_0}\Phi\left(\frac{1}{m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right).$$

Similarly,

$$\bar{n}_0\left(t_0, \frac{1}{1-m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) = \frac{1}{N_0}\Phi\left(\frac{1}{1-m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right).$$

Also, (A.9) implies

$$\begin{aligned}
\bar{n}_1(t,x) &= \frac{1}{2}\Bigg[\frac{1}{m}\bar{n}_0\left(t_0, \frac{1}{m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) \\
&\qquad + \frac{1}{1-m}\bar{n}_0\left(t_0, \frac{1}{1-m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right)\Bigg] \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right] \\
&= \frac{1}{2}\Bigg[\frac{1}{m}\frac{1}{N_0}\Phi\left(\frac{1}{m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right] \\
&\qquad + \frac{1}{1-m}\frac{1}{N_0}\Phi\left(\frac{1}{1-m}x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\Bigg] \\
&= \frac{1}{2}\left[\frac{1}{m}\bar{n}_0(t, \tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_0(t, \tfrac{1}{1-m}x)\right].
\end{aligned}$$

Thus, the claim holds for $i = 1$.

For the inductive step, we assume that the claim holds for $i = k$; i.e.,

$$2\bar{n}_k(t,x) = \frac{1}{m}\bar{n}_{k-1}(t, \tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_{k-1}(t, \tfrac{1}{1-m}x). \tag{A.11}$$

Substituting t_0 for t and ξ for x, this becomes

$$2\bar{n}_k(t_0,\xi) = \frac{1}{m}\bar{n}_{k-1}(t_0, \tfrac{1}{m}\xi) + \frac{1}{1-m}\bar{n}_{k-1}(t_0, \tfrac{1}{1-m}\xi). \tag{A.12}$$

If we let $\varphi(t) = \exp\left[\int_{t_0}^{t} v(u)\,du\right]$, then we can use (A.9) to write

$$\bar{n}_k(t,x) = \frac{1}{2}\left[\frac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}x\cdot\varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}x\cdot\varphi(t)\Big)\right]\cdot\varphi(t), \tag{A.13}$$

which is equivalent to

$$2\bar{n}_k(t,x) = \left[\frac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}x\cdot\varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}x\cdot\varphi(t)\Big)\right]\cdot\varphi(t). \tag{A.14}$$

Now, (A.11) and (A.14) imply

$$\left[\frac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}x\cdot\varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}x\cdot\varphi(t)\Big)\right]\cdot\varphi(t) = \frac{1}{m}\bar{n}_{k-1}(t, \tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_{k-1}(t, \tfrac{1}{1-m}x). \tag{A.15}$$

Next, we can use (A.9) to write

$$\bar{n}_{k+1}(t,x) = \frac{1}{2}\left[\frac{1}{m}\bar{n}_{k}\Big(t_0, \tfrac{1}{m}x\cdot\varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{k}\Big(t_0, \tfrac{1}{1-m}x\cdot\varphi(t)\Big)\right]\cdot\varphi(t),$$

which is equivalent to

$$2\bar{n}_{k+1}(t,x) = \left[\frac{1}{m}\bar{n}_{k}\Big(t_0, \tfrac{1}{m}x\cdot\varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{k}\Big(t_0, \tfrac{1}{1-m}x\cdot\varphi(t)\Big)\right]\cdot\varphi(t). \tag{A.16}$$

But (A.12) with $\xi = \frac{1}{m}x\cdot\varphi(t)$ implies

$$\bar{n}_k\Big(t_0, \tfrac{1}{m}x\cdot\varphi(t)\Big) = \frac{1}{2}\left[\frac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}\tfrac{1}{m}x\cdot\varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}\tfrac{1}{m}x\cdot\varphi(t)\Big)\right] \tag{A.17}$$

and (A.12) with $\xi = \frac{1}{1-m}x \cdot \varphi(t)$ implies

$$\bar{n}_k\Big(t_0, \tfrac{1}{1-m}x \cdot \varphi(t)\Big) = \frac{1}{2}\Bigg[\frac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}\tfrac{1}{1-m}x \cdot \varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}\tfrac{1}{1-m}x \cdot \varphi(t)\Big)\Bigg], \tag{A.18}$$

Substituting (A.17) and (A.18) into (A.16), we obtain

$$\begin{aligned} 2\bar{n}_{k+1}(t,x) &= \Bigg\{\frac{1}{m}\cdot\frac{1}{2}\Big[\tfrac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}\tfrac{1}{m}x \cdot \varphi(t)\Big) + \tfrac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}\tfrac{1}{m}x \cdot \varphi(t)\Big)\Big] \\ &\quad + \frac{1}{1-m}\cdot\frac{1}{2}\Big[\tfrac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}\tfrac{1}{1-m}x \cdot \varphi(t)\Big) + \tfrac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}\tfrac{1}{1-m}x \cdot \varphi(t)\Big)\Big]\Bigg\} \\ &\qquad \cdot \varphi(t) \\ &= \frac{1}{m}\cdot\frac{1}{2}\Big[\tfrac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}\tfrac{1}{m}x \cdot \varphi(t)\Big) + \tfrac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}\tfrac{1}{m}x \cdot \varphi(t)\Big)\Big] \cdot \varphi(t) \\ &\quad + \frac{1}{1-m}\cdot\frac{1}{2}\Big[\tfrac{1}{m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{m}\tfrac{1}{1-m}x \cdot \varphi(t)\Big) + \tfrac{1}{1-m}\bar{n}_{k-1}\Big(t_0, \tfrac{1}{1-m}\tfrac{1}{1-m}x \cdot \varphi(t)\Big)\Big] \cdot \varphi(t) \end{aligned} \tag{A.19}$$

If we use the identity from (A.15) with x replaced by $\frac{1}{m}x$, we can make a substitution in the first term of (A.19) that removes dependence on t_0. Similarly, using (A.15) with x replaced by $\frac{1}{1-m}x$, we can make a substitution in the second term of (A.19) that removes dependence on t_0. The resulting expression is

$$\begin{aligned} 2\bar{n}_{k+1}(t,x) &= \frac{1}{m}\cdot\frac{1}{2}\Big[\tfrac{1}{m}\bar{n}_{k-1}\Big(t, \tfrac{1}{m}\tfrac{1}{m}x\Big) + \tfrac{1}{1-m}\bar{n}_{k-1}\Big(t, \tfrac{1}{1-m}\tfrac{1}{m}x\Big)\Big] \\ &\quad + \frac{1}{1-m}\cdot\frac{1}{2}\Big[\tfrac{1}{m}\bar{n}_{k-1}\Big(t, \tfrac{1}{m}\tfrac{1}{1-m}x\Big) + \tfrac{1}{1-m}\bar{n}_{k-1}\Big(t, \tfrac{1}{1-m}\tfrac{1}{1-m}x\Big)\Big]. \end{aligned} \tag{A.20}$$

In order to make two final substitutions, note that replacing x with $\frac{1}{m}x$ in (A.11) yields

$$2\bar{n}_k(t, \tfrac{1}{m}x) = \tfrac{1}{m}\bar{n}_{k-1}(t, \tfrac{1}{m}\tfrac{1}{m}x) + \tfrac{1}{1-m}\bar{n}_{k-1}(t, \tfrac{1}{1-m}\tfrac{1}{m}x), \tag{A.21}$$

while replacing x with $\frac{1}{1-m}x$ in (A.11) yields

$$2\bar{n}_k(t, \tfrac{1}{1-m}x) = \tfrac{1}{m}\bar{n}_{k-1}(t, \tfrac{1}{m}\tfrac{1}{1-m}x) + \tfrac{1}{1-m}\bar{n}_{k-1}(t, \tfrac{1}{1-m}\tfrac{1}{1-m}x). \tag{A.22}$$

Substituting (A.21) and (A.22) into (A.20), we arrive at

$$\begin{aligned} 2\bar{n}_{k+1}(t,x) &= \frac{1}{m}\cdot\frac{1}{2}\Big[2\bar{n}_k(t,\tfrac{1}{m}x)\Big] + \frac{1}{1-m}\cdot\frac{1}{2}\Big[2\bar{n}_k(t,\tfrac{1}{1-m}x)\Big] \\ &= \frac{1}{m}\bar{n}_k(t,\tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_k(t,\tfrac{1}{1-m}x). \end{aligned}$$

Thus, we have demonstrated that if the claim holds for $i = k$, it must also hold for $i = k+1$. By the Principle of Mathematical Induction, the claim therefore holds for all $i \in \{1, 2, \ldots\}$. □

A.3 Relative Variation in Cell Counts Is Constant With Respect to Time

In the sections that follow, we consider populations (cultures) of cells which proliferate according the mathematical model described by (2.11) through (2.13) and (2.6). Working under the assumption that this model correctly describes cell numbers (counts), we show that two different measures of relative variation, percent difference and coefficient of variation, must remain constant in time.

A.3.1 Percent Difference Is Constant

Let $N(t)$ denote the total number of cells in a population at time t. Then

$$N(t) = \sum_{i=0}^{\infty} N_i(t),$$

where each $N_i(t)$ indicates the number of cells having completed i divisions at time t. We would like to show that two distinct cultures of cells (such as those cultures in two distinct wells) that are proliferating according to the same dynamics (or the same model parameters) will maintain the same "percent difference" in their cell numbers for all times t. So, for example, if

$$N^A(t) = \sum_{i=0}^{\infty} N_i^A(t)$$

and

$$N^B(t) = \sum_{i=0}^{\infty} N_i^B(t)$$

represent the total number of cells at time t in populations A and B, respectively, we want to demonstrate that

$$\frac{2(N^A(t_1) - N^B(t_1))}{N^A(t_1) + N^B(t_1)} = \frac{2(N^A(t_2) - N^B(t_2))}{N^A(t_2) + N^B(t_2)}$$

for arbitrary times $t_1, t_2 \in [t_0, \infty)$.

If we let $N_0^A(t_0) \equiv N_0^A$, then we can use (2.11) to write

$$\begin{aligned}
N^A(t) &= \sum_{i=0}^{\infty} N_i^A(t) \\
&= N_0^A(t) + N_1^A(t) + N_2^A(t) + \cdots \\
&= \left[N_0^A - \int_{t_0}^{t} (n_0^{div}(s) + n_0^{die}(s))\, ds \right] + \left[\int_{t_0}^{t} (2n_0^{div}(s) - n_1^{div}(s) - n_1^{div}(s))\, ds \right] \\
&\qquad + \left[\int_{t_0}^{t} (2n_1^{div}(s) - n_2^{div}(s) - n_2^{div}(s))\, ds \right] + \cdots \\
&= N_0^A + \int_{t_0}^{t} \Bigg(\left[n_0^{div,A}(s) - n_0^{die,A}(s) \right] + \left[n_1^{div,A}(s) - n_1^{die,A}(s) \right] \\
&\qquad\qquad + \left[n_2^{div,A}(s) - n_2^{die,A}(s) \right] + \cdots \Bigg)\, ds.
\end{aligned}$$

Now, observe that the expressions for $n_0^{div}(t)$ and $n_0^{die}(t)$ given in (2.12) and (2.13) both contain a constant factor of N_0 so that we can write

$$n_0^{div,A}(s) = N_0^A g_0^{div}(s)$$

and

$$n_0^{die,A}(s) = N_0^A g_0^{die}(s).$$

Similarly, through the recursive relationships indicated in (2.12) and (2.13), each $n_i^{div}(t)$ and each $n_i^{die}(t)$ contains a constant factor of N_0. Thus we can write

$$n_i^{div,A}(s) = N_0^A g_i^{div}(s)$$

and

$$n_i^{die,A}(s) = N_0^A g_i^{die}(s)$$

for each $i \geq 1$. Therefore, we have that

$$\begin{aligned} N^A(t) &= N_0^A + \int_{t_0}^{t} \Big(\Big[N_0^A g_0^{div}(s) - N_0^A g_0^{die}(s) \Big] + \Big[N_0^A g_1^{div}(s) - N_0^A g_1^{die}(s) \Big] \\ &\qquad + \Big[N_0^A g_2^{div}(s) - N_0^A g_2^{die}(s) \Big] + \cdots \Big) \, ds \\ &= N_0^A \left(1 + \int_{t_0}^{t} \sum_{i=0}^{\infty} \Big[g_i^{div}(s) - g_i^{die}(s) \Big] \, ds \right) \\ &= N_0^A g(t), \end{aligned}$$

where

$$g(t) = 1 + \int_{t_0}^{t} \sum_{i=0}^{\infty} \Big[g_i^{div}(s) - g_i^{die}(s) \Big] \, ds.$$

Similarly, it can be argued that

$$N^B(t) = N_0^B g(t).$$

Thus,

$$\begin{aligned} \frac{2(N^A(t) - N^B(t))}{N^A(t) + N^B(t)} &= \frac{2(N_0^A g(t) - N_0^B g(t))}{N_0^A g(t) + N_0^B g(t)} = \frac{2(N_0^A - N_0^B) g(t)}{(N_0^A + N_0^B) g(t)} \\ &= \frac{2(N_0^A - N_0^B)}{(N_0^A + N_0^B)}, \end{aligned}$$

for all t such that $g(t)$ is nonzero. (Note that $g(t) = 0$ implies that the number of cells in both cultures at time t is zero, in which case there is *no* difference in the numbers of cells in the two cultures.) So, assuming our model (2.14) for cell proliferation dynamics is correct, percent difference in the number of cells in two distinct cultures remains constant in time.

A.3.2 Coefficient of Variation Is Constant

It is also true that the "coefficient of variation" for numbers of cells in a sample of distinct cultures that is drawn from the set of all cultures proliferating according to the same dynamics should remain the same at all points in time. Recall that the coefficient of variation can be estimated as

$$\hat{c}_v = \frac{s}{\bar{x}},$$

where $\bar{x}$ and s are the mean and standard deviation for a sample of the "population" in question – in this case, the population would be all cultures of cells proliferating according to the same dynamics. If we have three cultures, call them A, B, and C, the sample mean population number

at time t is

$$\bar{x}(t) = \frac{N^A(t) + N^B(t) + N^C(t)}{3} = \frac{N_0^A + N_0^B + N_0^C}{3} g(t) = \bar{N}_0 g(t)$$

and the sample standard deviation in the population number is

$$\begin{aligned} s(t) &= \sqrt{\frac{1}{2}\Big((N^A(t) - \bar{x}(t))^2 + (N^B(t) - \bar{x}(t))^2 + (N^C(t) - \bar{x}(t))^2\Big)} \\ &= \sqrt{\frac{1}{2}\Big((N_0^A g(t) - \bar{N}_0 g(t))^2 + (N_0^B g(t) - \bar{N}_0 g(t))^2 + (N_0^C g(t) - \bar{N}_0 g(t))^2\Big)} \\ &= g(t)\sqrt{\frac{1}{2}\Big((N_0^A - \bar{N}_0)^2 + (N_0^B - \bar{N}_0)^2 + (N_0^C - \bar{N}_0)^2\Big)}. \end{aligned}$$

Thus,

$$\hat{c}_v(t) = \frac{s(t)}{\bar{x}(t)} = \frac{\sqrt{\frac{1}{2}\Big((N_0^A - \bar{N}_0)^2 + (N_0^B - \bar{N}_0)^2 + (N_0^C - \bar{N}_0)^2\Big)}}{\bar{N}_0} = \hat{c}_v(t_0)$$

for all t such that $g(t)$ is nonzero. So, assuming our model (2.14) for cell proliferation dynamics is correct, coefficient of variation in the number of cells in three distinct cultures remains constant in time. Here we have used a sample size of three since our experiments were performed in triplicate, but the proof is similar for any sample size.

Appendix B

Numerical Methods and Implementation Details

As discussed in [5], our numerical methods for solving the label-structured cyton model for cell densities (2.14) are an extension (accounting for the incorporation of the cyton model) of the methods originally proposed by Hasenauer et al. in [27]. Because the solutions $\{n_i(t,x)\}$ of (2.14) are factorable (cf. Proposition 2.2 and Proposition 5.2), one can compute the cell numbers $\{N_i(t)\}$ and the distributions of CFSE FI $\{\bar{n}_i(t,x)\}$ independently. Then, the solutions of (2.14) can be computed as $n_i(t,x) = N_i(t)\bar{n}_i(t,x)$ for all i, and the autofluorescence-adjusted structured densities $\{\tilde{n}_i(t,\tilde{x})\}$ can be computed using the convolution integral (2.10).

In this appendix, we discuss the numerical methods used for computing the CFSE FI distributions $\{\bar{n}_i(t,x)\}$, the cell numbers $\{N_i(t)\}$, the initial condition $\Phi(x)$, and the total structured density $\tilde{n}(t,\tilde{x})$. We also provide details of our inverse (parameter estimation) problem implementation.

B.1 Computation of CFSE FI Distributions $\{\bar{n}_i(t,x)\}$

We consider two cases in turn. In the first, all cell divisions are assumed to be symmetric in the sense that each mother cell produces two daughter cells that both receive one half of the CFSE that was present in the mother cell. In the second, we assume divisions can be asymmetric, with the two daughter cells receiving proportions m and $1-m$, respectively, of the CFSE that was present in the mother cell. In the latter case, $m \in (0, \frac{1}{2}]$ is assumed to be a constant for the population of cells in question.

B.1.1 Symmetric Cell Division

For the case in which all cell divisions are assumed to be symmetric, we demonstrate in Appendix A.1 that

$$\bar{n}_i(t,x) = \frac{2^i}{N_0}\Phi\left(2^i x \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]\right) \cdot \exp\left[\int_{t_0}^{t} v(u)\,du\right]$$

is the solution of (2.7) and (2.8) for each $i \in \{0,1,2,\ldots\}$. Letting $\varphi(t) = \exp\left[\int_{t_0}^{t} v(u)\,du\right]$ to simplify notation, we can write this as

$$\bar{n}_i(t,x) = \frac{2^i\varphi(t)}{N_0}\Phi\Big(2^i x \cdot \varphi(t)\Big) \tag{B.1}$$

for each $i \in \{0,1,2,\ldots\}$.

Now, following the work of Hasenauer et al. [27], we make the further assumption that the initial CFSE FI distribution is lognormal with parameters μ_0 and σ_0^2; i.e.,

$$\bar{n}_0(t_0,x) = \mathrm{logn}(x;\mu_0,\sigma_0^2) = \frac{1}{x\sigma_0\sqrt{2\pi}} \cdot \exp\left[\frac{-(\log x - \mu_0)^2}{2\sigma_0^2}\right].$$

(We will extend our argument to a considerably less restrictive assumption in a moment.) Then from (2.8) we have the relation

$$\frac{\Phi(x)}{N_0} = \bar{n}_0(t_0,x) = \frac{1}{x\sigma_0\sqrt{2\pi}} \cdot \exp\left[\frac{-(\log x - \mu_0)^2}{2\sigma_0^2}\right] \tag{B.2}$$

for $x > 0$. If we replace x by $2^i x \cdot \varphi(t)$ in (B.2) and substitute the resulting expression into

(B.1), we obtain

$$\begin{aligned}
\bar{n}_i(t,x) &= \frac{2^i\varphi(t)}{\Big(2^i x\cdot\varphi(t)\Big)\sigma_0\sqrt{2\pi}}\cdot\exp\left[\frac{-\Big(\log\Big(2^i x\cdot\varphi(t)\Big)-\mu_0\Big)^2}{2\sigma_0^2}\right] \\
&= \frac{1}{x\sigma_0\sqrt{2\pi}}\cdot\exp\left[\frac{-\Big(\log x + i\log 2+\log\varphi(t)-\mu_0\Big)^2}{2\sigma_0^2}\right] \\
&= \frac{1}{x\sigma_0\sqrt{2\pi}}\cdot\exp\left[\frac{-\Big(\log x-\Big(-i\log 2-\log\varphi(t)+\mu_0\Big)\Big)^2}{2\sigma_0^2}\right] \\
&= \frac{1}{x\sigma_0\sqrt{2\pi}}\cdot\exp\left[\frac{-\Big(\log x-\Big(-i\log 2-\int_{t_0}^t v(u)\,du+\mu_0\Big)\Big)^2}{2\sigma_0^2}\right] \\
&= \frac{1}{x\sigma_0\sqrt{2\pi}}\cdot\exp\left[\frac{-\Big(\log x-\mu_i(t)\Big)^2}{2\sigma_0^2}\right] \\
&= \mathrm{logn}(x;\mu_i(t),\sigma_0^2), \qquad \text{(B.3)}
\end{aligned}$$

where

$$\mu_i(t) = -i\log 2 - \int_{t_0}^t v(u)\,du + \mu_0.$$

In other words, if the initial CFSE FI distribution (at time t_0) is lognormal with parameters μ_0 and σ_0^2, then the CFSE FI distribution for any generation $i \geq 0$ will be lognormal at time $t > t_0$ with parameters $\mu_i(t)$ and σ_0^2.

Still following the work of Hasenauer et al. [27], we next consider a less restrictive assumption on the form of the initial CFSE FI distribution. Suppose that the initial structured density can be written using a linear (convex) combination of lognormal pdfs as

$$\Phi(x) = N_0\sum_{k=1}^{k_{\max}} a_k \mathrm{logn}(x;\mu^k,(\sigma^k)^2),$$

where $a_k \geq 0$ for $k \in \{1,\ldots,k_{\max}\}$ and $\sum_{k=1}^{k_{\max}} a_k = 1$. This assumption is not overly restrictive, and in practice we find that the initial structured densities observed in CFSE-based flow cytometry experiments can be well-approximated by such a series with $k_{\max} = 3$. (More will be said about our methods for determining $\{a_k\}$, $\{\mu^k\}$, and $\{\sigma^k\}$ in Section B.3.) Then by the

principle of superposition and (B.3),

$$\bar{n}_i(t,x) = \sum_{k=1}^{k_{\max}} a_k \text{logn}(x; \mu_i^k(t), (\sigma^k)^2), \tag{B.4}$$

where

$$\mu_i^k(t) = -i \log 2 - \int_{t_0}^{t} v(u)\, du + \mu^k \tag{B.5}$$

for all i and $k \in \{1, \ldots, k_{\max}\}$. For the specific models we employ in this dissertation, $v(u) = c$, where $c > 0$ is some constant (cf. Section 2.5); therefore, the integrals in (B.5) can be trivially evaluated as $c(t - t_0)$, and we can compute $\{\bar{n}_i(t,x)\}$ for any values of $i \in \{0, 1, \ldots\}$, $t > t_0$, and $x > 0$ that we choose. As discussed in Section 2.1, we typically only consider those cells that have divided up to $i_{\max} = 16$ times, so computations are only performed for $i \in \{0, \ldots, 16\}$.

B.1.2 Asymmetric Cell Division

In the case of asymmetric cell division, we unfortunately cannot express $\bar{n}_i(t,x)$ as a linear combination of lognormal pdfs for arbitrary t. Instead, we will use recursion to define $\bar{n}_i(t,x)$. We first consider the situation when $t = t_0$, and then consider the more general case $t > t_0$. Note that, just as in the symmetric cell division case, we assume that the initial CFSE FI distribution can be approximated by a linear combination of lognormal pdfs; i.e.,

$$\bar{n}_0(t_0,x) = \sum_{k=1}^{k_{\max}} a_k \text{logn}(x; \mu^k, (\sigma^k)^2). \tag{B.6}$$

for some $\{a_k\}$, $\{\mu^k\}$, and $\{\sigma^k\}$. (Again, refer to Section B.3 for a discussion of how these initial condition parameters may be determined.)

For $t = t_0$, Lemma A.2 implies

$$\bar{n}_i(t_0,x) = \frac{1}{2}\left[\frac{1}{m}\bar{n}_{i-1}(t_0, \tfrac{1}{m}x) + \frac{1}{1-m}\bar{n}_{i-1}(t_0, \tfrac{1}{1-m}x)\right]. \tag{B.7}$$

Therefore, we can determine the value of $\bar{n}_i(t_0,x)$ for any generation i at any FI x through a recursion on i using (B.6) and (B.7).

For the case $t > t_0$, we use (A.10) and the function φ defined in the proof of Lemma A.2 along with the fact (cf. (5.2)) that

$$\frac{\Phi(\xi)}{N_0} = \bar{n}_0(t_0, \xi)$$

to write

$$\bar{n}_0(t,x) = \bar{n}_0\big(t_0, x \cdot \varphi(t)\big) \cdot \varphi(t).$$

Also we can rewrite (A.13) with the index k replaced by i to obtain

$$\bar{n}_i(t,x) = \frac{1}{2}\left[\frac{1}{m}\bar{n}_{i-1}\Big(t_0, \tfrac{1}{m}x\cdot\varphi(t)\Big) + \frac{1}{1-m}\bar{n}_{i-1}\Big(t_0, \tfrac{1}{1-m}x\cdot\varphi(t)\Big)\right]\cdot\varphi(t).$$

The two foregoing equations allow us to compute the value of $\bar{n}_i(t,x)$ for any generation i, any time $t > t_0$, and any FI x using evaluations of $\bar{n}_i(t_0, \gamma(t,x))$, where $\gamma(t,x)$ is a known value given by either $\frac{1}{m}x\cdot\varphi(t)$ or $\frac{1}{1-m}x\cdot\varphi(t)$. Thus, for any $t > t_0$, we can translate the density $\bar{n}_i(t,x)$ such that only evaluations of $\bar{n}_i$ in which $t = t_0$ are needed. These evaluations can be performed using recursion as demonstrated in the previous paragraph. Note that because of the compounding expense of recursive function calls, we choose to only consider cells that have divided up to $i_{\max} = 8$ times. That is, we only perform computations for $i \in \{0,\ldots,8\}$. We find that in a typical five-day CFSE flow cytometry experiment, the number of cells that divide more than 8 times is negligible. Therefore, this simplification does not adversely impact the ability of our model to capture the aggregate behavior of a population of proliferating cells. Further arguments in support of this choice for $i_{\max}$ are provided in Section 5.5.

B.2 Computation of Cell Numbers $\{N_i(t)\}$

We next describe our methods for computing solutions $\{N_i(t)\}$ to the cyton model (2.11) subject to the initial conditions (2.6). We assume that ϕ_i and ψ_i, which represent pdfs for time until division and time until death, respectively, of cells in generation i, are known functions of time for all $i \in \{0,1,\ldots\}$. In order to evaluate expressions involved in the cyton model, we need to compute and store values of $\phi_0(t)$, $\psi_0(t)$, $\int_{t_0}^t \phi_0(s)\,ds$, and $\int_{t_0}^t \psi_0(s)\,ds$ for various values of t. We also need to compute and store values of $\phi_i(t)$, $\psi_i(t)$, $\int_0^t \phi_i(s)\,ds$, and $\int_0^t \psi_i(s)\,ds$, where $i \geq 1$, for various values of t. We describe our methods for completing each of these tasks in the following sections.

B.2.1 Numerical Evaluation of Expressions Involving ϕ_0 and ψ_0

To begin, we consider the cyton functions ϕ_0 and ψ_0 associated with generation $i = 0$. Recall that we interpret ϕ_0 and ψ_0 as the pdfs for time until division and time until death, respectively, for undivided cells in the initial seed population (cf. Section 2.3), but recall also that *we only consider in our models those cells from the initial population that have neither died nor divided as of time* $t_0 \approx 24$ *hours* (cf. Section 2.5). That is, since we have no information about the behavior of cells in the time interval $[0, t_0]$, we assume that *nothing* happens to the *cells under consideration* during this time. For this reason, we must take care to ensure that $\phi_0(t)$ and $\psi_0(t)$ are only nonzero for values of $t > t_0$. More will be said about this in a moment.

Now, if we consider an experiment in which the first observation occurs at time t_0 and the final observation occurs at time t_f, we can use a time step size h (selected such that it divides $t_f - t_0$) to partition the the interval $[t_0, t_f]$ using the nodal points

$$\{t_j\} = \big\{t_0 + jh : j \in \{0, 1, \ldots, N_t\}\big\},$$

where $N_t = \frac{t_f - t_0}{h}$. We can then compute the values $\{\phi_0(t_j)\}$ and $\{\psi_0(t_j)\}$ and store them in two vectors of length $N_t + 1$.

Next, we compute the values $\{\int_{t_0}^{t_j} \phi_0(s)\, ds\}$ and $\{\int_{t_0}^{t_j} \psi_0(s)\, ds\}$ and store them in two additional vectors of length $N_t + 1$. The integrations involved can be efficiently carried out using two-point Gauss-Legendre quadrature [40] on each subinterval of length h. We use $h = 0.25$ for all of our computational purposes, as we have found that this value of h yields sufficiently high precision on the values of all numerically evaluated integrals.

Finally, we must say a word about the truncation and scaling of the functions ϕ_0 and ψ_0, as well as the scaling of the four stored vectors $\{\phi_0(t_j)\}$, $\{\psi_0(t_j)\}$, $\{\int_{t_0}^{t_j} \phi_0(s)\, ds\}$, and $\{\int_{t_0}^{t_j} \psi_0(s)\, ds\}$, which we have described in the preceding paragraphs. Since we make the assumption that ϕ_0 and ψ_0 are *lognormal* pdfs (cf. Section 2.5), which only have support on the interval $(0, \infty)$, we can define the modified pdfs $\tilde{\phi}_0$ and $\tilde{\psi}_0$ by

$$\tilde{\phi}_0(t) = \begin{cases} \frac{\phi_0(t)}{1 - \int_0^{t_0} \phi_0(s)\, ds} & \text{for } t \in (t_0, \infty), \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\tilde{\psi}_0(t) = \begin{cases} \frac{\psi_0(t)}{1 - \int_0^{t_0} \psi_0(s)\, ds} & \text{for } t \in (t_0, \infty), \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, these truncated and scaled functions are also pdfs and, furthermore, they have support only on the interval (t_0, ∞) as desired. In order to compute the integrals in the denominators of the above expressions, we select a step size h (which divides t_0) and partition the time interval $[0, t_0]$ using the $\frac{t_0}{h} + 1$ nodes $\{0, h, \ldots, t_0\}$. For our purposes, we once again take $h = 0.25$. As described previously, the integrations can then be efficiently carried out using two-point Gauss-Legendre quadrature on each subinterval of length h. The values in the stored vectors $\{\phi_0(t_j)\}$ and $\{\int_{t_0}^{t_j} \phi_0(s)\, ds\}$ can now be divided by $1 - \int_0^{t_0} \phi_0(s)\, ds$ to produce the vectors $\{\tilde{\phi}_0(t_j)\}$ and $\{\int_{t_0}^{t_j} \tilde{\phi}_0(s)\, ds\}$. In the event that $1 - \int_0^{t_0} \phi_0(s)\, ds$ is numerically zero, we set all the values in both vectors to zero. Similarly, the values in the stored vectors $\{\psi_0(t_j)\}$ and $\{\int_{t_0}^{t_j} \psi_0(s)\, ds\}$ can be scaled appropriately to produce the vectors $\{\tilde{\psi}_0(t_j)\}$ and $\{\int_{t_0}^{t_j} \tilde{\psi}_0(s)\, ds\}$. Hereafter, we will simply refer to the modified functions (and the corresponding vectors) using the symbols ϕ_0

and ψ_0 without tildes, but it should be understood that the functions have been truncated and scaled as described above.

As a result of the computational work described in this section, we now have four vectors of length $N_t + 1$:

$$\{\phi_0(t_j)\} = \Big[\phi_0(t_0), \phi_0(t_0 + h), \ldots, \phi_0(t_f)\Big],$$

$$\{\psi_0(t_j)\} = \Big[\psi_0(t_0), \psi_0(t_0 + h), \ldots, \psi_0(t_f)\Big],$$

$$\Big\{ \textstyle\int_{t_0}^{t_j} \phi_0(s)\, ds \Big\} \approx \Big[\textstyle\int_{t_0}^{t_0} \phi_0(s)\, ds, \int_{t_0}^{t_0+h} \phi_0(s)\, ds, \ldots, \int_{t_0}^{t_f} \phi_0(s)\, ds \Big],$$

and

$$\Big\{ \textstyle\int_{t_0}^{t_j} \psi_0(s)\, ds \Big\} \approx \Big[\textstyle\int_{t_0}^{t_0} \psi_0(s)\, ds, \int_{t_0}^{t_0+h} \psi_0(s)\, ds, \ldots, \int_{t_0}^{t_f} \psi_0(s)\, ds \Big],$$

where "$\approx$" indicates that the integrals in question have been approximated using cumulative sums of integrals on subintervals of length h that were computed using Gauss-Legendre quadrature.

B.2.2 Numerical Evaluation of Expressions Involving ϕ_i and ψ_i for $i \geq 1$

Next, we consider expressions involving ϕ_i and ψ_i for generations $i \geq 1$. Recall from Section 2.5 that we assume $\phi_i(t) = \phi_k(t)$ and $\psi_i(t) = \psi_k(t)$ for all $i, k \in \{1, 2, \ldots\}$. Therefore, for computational purposes we really only need to consider ϕ_1 and ψ_1.

Expressions in the cyton model require us to evaluate $\phi_i(t) = \phi_1(t)$ and $\psi_i(t) = \psi_1(t)$ (where $i \geq 1$) at values of $t \in [0, t_f - t_0]$, so we create a partition $\{\tilde{t}_j\} = \{t_j - t_0\} = \{0, h, 2h, \ldots, t_f - t_0\}$ for this time interval using the same step size h defined previously. Noting that this partition consists of $N_t + 1$ nodal points, we can compute the values $\{\phi_1(\tilde{t}_j)\}$ and $\{\psi_1(\tilde{t}_j)\}$ and store them in two vectors of length $N_t + 1$.

Next, we need to compute the values $\{\int_0^{\tilde{t}_j} \phi_1(s)\, ds\}$ and $\{\int_0^{\tilde{t}_j} \psi_1(s)\, ds\}$ and store them in two additional vectors of length $N_t + 1$. As discussed in the previous section, the necessary values can be computed to high precision by using the values stored in $\{\phi_1(\tilde{t}_j)\}$ and $\{\psi_1(\tilde{t}_j\}$ to compute two-point Gauss-Legendre quadratures on each subinterval of length h.

As a result of the computational work described in this section, we now have an additional four vectors of length $N_t + 1$:

$$\{\phi_1(\tilde{t}_j)\} = \Big[\phi_1(0), \phi_1(h), \ldots, \phi_1(t_f - t_0)\Big],$$

$$\{\psi_1(\tilde{t}_j)\} = \Big[\psi_1(0), \psi_1(h), \ldots, \psi_1(t_f - t_0)\Big],$$

$$\Big\{ \textstyle\int_0^{\tilde{t}_j} \phi_1(s)\, ds \Big\} \approx \Big[\textstyle\int_0^{0} \phi_1(s)\, ds, \int_0^{h} \phi_1(s)\, ds, \ldots, \int_0^{t_f - t_0} \phi_1(s)\, ds \Big],$$

and

$$\left\{ \int_0^{\tilde{t}_j} \psi_1(s)\, ds \right\} \approx \left[\int_0^0 \psi_1(s)\, ds, \int_0^h \psi_1(s)\, ds, \ldots, \int_0^{t_f - t_0} \psi_1(s)\, ds \right].$$

Once again, "$\approx$" indicates that the integrals in question have been approximated using cumulative sums of integrals on subintervals of length h that were computed using Gauss-Legendre quadrature.

B.2.3 Numerical Evaluation of $\{n_i^{div}(t_j)\}$ and $\{n_i^{die}(t_j)\}$

We now explain how to numerically evaluate the cyton division and death rates $\{n_i^{div}(t_j)\}$ and $\{n_i^{die}(t_j)\}$. For the case in which $i = 0$, we use the (scaled) vectors $\{\int_{t_0}^{t_j} \psi_0(s)\, ds\}$ and $\{\phi_0(t_j)\}$ to compute the values $\{n_0^{div}(t_j)\}$ according to (2.12) and then store them in a vector of length $N_t + 1$. Note that the work required to compute the vector $\{n_0^{div}(t_j)\}$ is dominated by a single element-wise multiplication of the two vectors $\{1 - \int_{t_0}^{t_j} \psi_0(s)\, ds\}$ and $\{\phi_0(t_j)\}$. One of these vectors has already been computed and stored, and the other can be easily obtained by changing the signs of the elements in a previously stored vector and adding 1 to each of them. Similarly, we can use the vectors $\{\int_{t_0}^{t_j} \phi_0(s)\, ds\}$ and $\{\psi_0(t_j)\}$ to compute the values $\{n_0^{die}(t_j)\}$ according to (2.13) and store them in a vector of length $N_t + 1$. The work of computing the values for the vector $\{n_0^{die}(t_j)\}$ is also dominated by a single element-wise multiplication of two vectors.

For the case in which $i \geq 1$, we need values for expressions of the form $\phi_i(t_j - t_k)$ and $\psi_i(t_j - t_k)$, where $j \geq k \geq 0$. Note, however, that because the nodes $\{t_j\} = \{\tilde{t}_j + t_0\}$ are evenly spaced and because the cytons for $i \geq 1$ are equivalent, we have the relations

$$\phi_i(t_j - t_k) = \phi_i(\tilde{t}_j - \tilde{t}_k) = \phi_i(\tilde{t}_{j-k}) = \phi_1(\tilde{t}_{j-k})$$

and

$$\psi_i(t_j - t_k) = \psi_i(\tilde{t}_j - \tilde{t}_k) = \psi_i(\tilde{t}_{j-k}) = \psi_1(\tilde{t}_{j-k})$$

for all $i \geq 1$ and all $j \geq k \geq 0$. Similarly,

$$\int_0^{t_j - t_k} \phi_i(\xi)\, d\xi = \int_0^{\tilde{t}_j - \tilde{t}_k} \phi_i(\xi)\, d\xi = \int_0^{\tilde{t}_{j-k}} \phi_i(\xi)\, d\xi = \int_0^{\tilde{t}_{j-k}} \phi_1(\xi)\, d\xi$$

and

$$\int_0^{t_j - t_k} \psi_i(\xi)\, d\xi = \int_0^{\tilde{t}_j - \tilde{t}_k} \psi_i(\xi)\, d\xi = \int_0^{\tilde{t}_{j-k}} \psi_i(\xi)\, d\xi = \int_0^{\tilde{t}_{j-k}} \psi_1(\xi)\, d\xi$$

for all $i \geq 1$ and all $j \geq k \geq 0$. That is, all the values for the expressions of the form $\phi_i(t_j - t_k)$, $\psi_i(t_j - t_k)$, $\int_0^{t_j - t_k} \phi_i(\xi)\, d\xi$, and $\int_0^{t_j - t_k} \psi_i(\xi)\, d\xi$, which will be needed to compute $\{n_i^{div}(t_j)\}$ and $\{n_i^{die}(t_j)\}$, have already been computed and stored!

Now, suppose we have already computed $\{n_{i-1}^{div}(t_j)\}_{j=0}^{N_t}$ and that we wish to compute $\{n_i^{div}(t_j)\}_{j=0}^{N_t}$. From (2.12), we have that

$$n_i^{div}(t_j) = 2F_i \int_{t_0}^{t_j} n_{i-1}^{div}(s)\Big(1 - \int_0^{t_j - s} \psi_i(\xi)\, d\xi\Big)\phi_i(t_j - s)\, ds$$

for any fixed value of j. Note that the integrand of the outer integral in this expression can be written

$$f_i^{div}(t_j, s) = n_{i-1}^{div}(s)\Big(1 - \int_0^{t_j - s} \psi_i(\xi)\, d\xi\Big)\phi_i(t_j - s).$$

Therefore, we can approximate the outer integral using a composite trapezoid quadrature if we can obtain the values

$$f_i^{div}(t_j, t_k) = n_{i-1}^{div}(t_k)\Big(1 - \int_0^{t_j - t_k} \psi_i(\xi)\, d\xi\Big)\phi_i(t_j - t_k)$$

for $k \in \{0, 1, \ldots, j\}$. But these values can be computed easily using an element-wise multiplication of three vectors of length $j + 1$ since

$$\begin{aligned}
&\Big\{n_i^{div}(t_k)\Big\}_{k=0}^{j}, \\
&\Big\{\textstyle\int_0^{t_j - t_k} \psi_i(\xi)\, d\xi\Big\}_{k=0}^{j} = \Big\{\textstyle\int_0^{\tilde{t}_{j-k}} \psi_1(\xi)\, d\xi\Big\}_{k=0}^{j}, \text{ and} \\
&\Big\{\phi_i(t_j - t_k)\Big\}_{k=0}^{j} = \Big\{\phi_1(\tilde{t}_{j-k})\Big\}_{k=0}^{j}
\end{aligned}$$

have already been computed.

A similar approach can be used to obtain $\{n_i^{die}(t_j)\}$ as follows. Again suppose that we have already computed $\{n_{i-1}^{div}(t_j)\}_{j=0}^{N_t}$, but this time consider that we wish to compute $\{n_i^{die}(t_j)\}_{j=0}^{N_t}$. From (2.13), we have that

$$n_i^{die}(t_j) = 2\int_{t_0}^{t_j} n_{i-1}^{div}(s)\left(1 - F_i \int_0^{t_j - s} \phi_i(\xi)\, d\xi\right)\psi_i(t_j - s)\, ds$$

for any fixed value of j. Note that the integrand of the outer integral in this expression can be written

$$f_i^{div}(t_j, s) = n_{i-1}^{div}(s)\left(1 - F_i \int_0^{t_j - s} \phi_i(\xi)\, d\xi\right)\psi_i(t_j - s).$$

Therefore, we can approximate the outer integral using a composite trapezoid quadrature if we can obtain the values

$$f_i^{div}(t_j, t_k) = n_{i-1}^{div}(t_k)\left(1 - F_i \int_0^{t_j - t_k} \phi_i(\xi)\, d\xi\right)\psi_i(t_j - t_k)$$

for $k \in \{0, 1, \ldots, j\}$. But these values can be computed easily using an element-wise multiplication of three vectors of length $j+1$ since

$$\begin{aligned}
&\left\{n_i^{div}(t_k)\right\}_{k=0}^{j}, \\
&\left\{\int_0^{t_j - t_k} \phi_i(\xi)\, d\xi\right\}_{k=0}^{j} = \left\{\int_0^{\tilde{t}_{j-k}} \phi_1(\xi)\, d\xi\right\}_{k=0}^{j}, \text{ and} \\
&\left\{\psi_i(t_j - t_k)\right\}_{k=0}^{j} = \left\{\psi_1(\tilde{t}_{j-k})\right\}_{k=0}^{j}
\end{aligned}$$

have already been computed.

As a result of the computational work described in this section, we now have an additional $2(i_{\max}+1)$ vectors of length $N_t + 1$:

$$\begin{aligned}
\{n_0^{div}(t_j)\} &= \left[n_0^{div}(t_0), n_0^{div}(t_0+h), \ldots, n_0^{div}(t_f)\right], \\
\{n_1^{div}(t_j)\} &= \left[n_1^{div}(t_0), n_1^{div}(t_0+h), \ldots, n_1^{div}(t_f)\right], \\
&\vdots \\
\{n_{i_{\max}+1}^{div}(t_j)\} &= \left[n_{i_{\max}+1}^{div}(t_0), n_{i_{\max}+1}^{div}(t_0+h), \ldots, n_{i_{\max}+1}^{div}(t_f)\right], \\
\{n_0^{die}(t_j)\} &= \left[n_0^{die}(t_0), n_0^{die}(t_0+h), \ldots, n_0^{die}(t_f)\right], \\
\{n_1^{die}(t_j)\} &= \left[n_1^{die}(t_0), n_1^{die}(t_0+h), \ldots, n_1^{die}(t_f)\right], \\
&\vdots \\
\{n_{i_{\max}+1}^{die}(t_j)\} &= \left[n_{i_{\max}+1}^{die}(t_0), n_{i_{\max}+1}^{die}(t_0+h), \ldots, n_{i_{\max}+1}^{die}(t_f)\right].
\end{aligned}$$

B.2.4 Numerical Evaluation of $\{N_i(t_j)\}$

Assuming we have the initial condition $N_0 = N_0(t_0)$, we can now compute $N_i(t_j)$ for $i \in \{0, 1, \ldots, i_{\max}\}$ and $j \in \{0, 1, \ldots, N_t\}$ using (2.11). (We discuss how to obtain the initial condition in Section B.3.) Our approach is to use the values stored in the vectors listed at the end of the previous section to approximate the integrals in (2.11) using composite trapezoid quadrature. Thus, we have shown how to obtain numerical solutions $\{N_i(t)\}_{i=0}^{i_{\max}}$ on the grid $\{t_j\} = \{t_0, t_0 + h, \ldots, t_f\}$.

B.3 Approximation of Initial Conditions

For a set of time series data such as that described in Section 1.2, we use summary histogram data from the first time point (i.e., the "Day 1" data) to construct an approximation of the initial CFSE FI distribution. The idea is to select parameters N_0, $k_{\max}$, $\{a_k\}$, $\{\mu^k\}$, and $\{\sigma^k\}$

such that the convex combination of lognormal pdfs given by

$$\Phi(x) = N_0 \sum_{k=1}^{k_{\max}} a_k \text{logn}(x; \mu^k, (\sigma^k)^2),$$

closely matches the actual summary histogram data for CFSE FI on Day 1. Note that the terminology "convex combination" implies that

$$a_k \geq 0 \text{ for } k \in \{1, \ldots, k_{\max}\} \quad \text{and} \quad \sum_{k=1}^{k_{\max}} a_k = 1. \tag{B.8}$$

We assume that we start with Day 1 summary histogram data in the form of two vectors x^{data} and y^{data}. The vector x^{data} has length $N^{\text{data}} + 1$ and its elements are the boundaries for the histogram bins (in units of FI). The vector y^{data} has length N^{data} and each of its elements gives the number of cells observed in the corresponding bin. To begin, we set

$$N'_{0,\text{init}} = \sum_{i=1}^{N^{\text{data}}} y_i^{\text{data}},$$

which is the true total number of cells represented in the data. We also compute the mean and variance of the FI data as

$$x^{\text{mean}} = \frac{1}{N'_{0,\text{init}}} \sum_{i=1}^{N^{\text{data}}} x_i^{\text{data}} y_i^{\text{data}}$$

and

$$x^{\text{var}} = \frac{1}{N'_{0,\text{init}}} \sum_{i=1}^{N^{\text{data}}} \left(x_i^{\text{data}} - \mu\right)^2 y_i^{\text{data}}.$$

Since the lognormal distribution with this mean and variance would have parameters

$$\mu'_{\text{init}} = \log\left(x^{\text{mean}}\right) - \frac{1}{2} \log\left(1 + \frac{x^{\text{var}}}{(x^{\text{mean}})^2}\right)$$

and

$$(\sigma'_{\text{init}})^2 = \log\left(1 + \frac{x^{\text{var}}}{(x^{\text{mean}})^2}\right),$$

we use these values as an initial iterate in our search for an optimal lognormal fit.

If we assume the initial structured density $\Phi(x)$ can be modeled by a single lognormal distribution with parameters μ and σ^2 that is scaled by a factor of N_0, the number of cells in

the bin with left bound x_i^{data} can be approximated as

$$N_0 \int_{x_i^{\text{data}}}^{x_{i+1}^{\text{data}}} \text{logn}(x; \mu, \sigma^2)\, dx.$$

Therefore we first seek a parameter vector $\theta' = (N_0', \mu', \sigma')$ that will minimize

$$J_1(\theta) = \sum_{i=1}^{N^{\text{data}}} \left(y_i^{\text{data}} - N_0 \int_{x_i^{\text{data}}}^{x_{i+1}^{\text{data}}} \text{logn}(x; \mu, \sigma^2)\, dx \right)^2,$$

where $\theta = (N_0, \mu, \sigma)$. Using an initial iterate $\theta'_{\text{init}} = (N'_{0,\text{init}}, \mu'_{\text{init}}, \sigma'_{\text{init}})$, we obtain such an optimal parameter vector θ' using the MATLAB routine `lsqnonlin`.

Next, we seek a better approximation for $\Phi(x)$ by considering a convex combination of *three* lognormal distributions. That is, we set $k_{\max} = 3$ and attempt to find an optimal parameter vector $\vartheta = (N_0, \{a_k\}, \{\mu^k\}, \{\sigma^k\})$. This is accomplished through an iterative scheme in which we alternate between estimating optimal values for $(N_0, \{a_k\})$ while fixing $(\{\mu^k\}, \{\sigma^k\})$ and estimating optimal values for $(N_0, \{\mu^k\}, \{\sigma^k\})$ while fixing $\{a_k\}$. The optimization scheme is described in Algorithm B.3.1, and relevant cost function is given by

$$J_2(\vartheta) = \sum_{i=1}^{N^{\text{data}}} \left(y_i^{\text{data}} - N_0 \sum_{k=1}^{k_{\max}} a_k \int_{x_i^{\text{data}}}^{x_{i+1}^{\text{data}}} \text{logn}(x; \mu^k, (\sigma^k)^2)\, dx \right)^2.$$

We implement the minimizations in Steps 6 and 7 of this algorithm using the MATLAB routine `fmincon`, which allows for optimization subject to the constraints given in (B.8).

B.4 Computation of Structured Density $\tilde{n}(t, \tilde{x})$

We define

$$n(t, x) = \sum_{i=0}^{i_{\max}} n_i(t, x)$$

to be the total (including *all* generations) structured density (in cells per unit FI), where $\{n_i(t, x)\}$ are the generation-indexed structured densities that satisfy (2.14). We discuss the different methods used in the cases of symmetric division and asymmetric division in the following two sections.

Algorithm B.3.1 Initial Condition Approximation Procedure

1. Compute the mean and standard deviation of the lognormal distribution with parameters μ' and $(\sigma')^2$ as
$$\mathcal{M} = e^{\mu' + (\sigma')^2/2}$$
and
$$\mathcal{S} = \mathcal{M} \cdot \sqrt{e^{(\sigma')^2} - 1}.$$
(Compare these formulae with (4.6) and (4.7).)

2. Create vectors $\vec{\mathcal{M}} = (\mathcal{M}, \mathcal{M}, 1.5\mathcal{M})$ and $\vec{\mathcal{S}} = (\mathcal{S}, 0.5\mathcal{S}, 3\mathcal{S})$ containing the means and standard deviations for three lognormal distributions.

3. Convert each mean and standard deviation pair into a pair of lognormal distribution parameters by setting
$$\mu^k_{\text{init}} = \log(\mathcal{M}_k) - \frac{1}{2}\log\left(1 + \frac{\mathcal{S}_k^2}{\mathcal{M}_k^2}\right)$$
and
$$\sigma^k_{\text{init}} = \sqrt{\log\left(1 + \frac{\mathcal{S}_k^2}{\mathcal{M}_k^2}\right)}$$
for $k \in \{1, 2, 3\}$, where $\mathcal{M}_k$ and $\mathcal{S}_k$ are the kth entries of $\vec{\mathcal{M}}$ and $\vec{\mathcal{S}}$, respectively. (Compare these formulae with (4.8) and (4.9).)

4. Set $\vartheta^{(0)} = (\vartheta_1^{(0)}, \vartheta_2^{(0)}, \vartheta_3^{(0)}, \vartheta_4^{(0)}) = (N_0', \{0.5, 0.35, 0.15\}, \{\mu^k_{\text{init}}\}, \{\sigma^k_{\text{init}}\})$ as the initial iterate.

5. Initialize the iteration counter ℓ with the value 1.

6. Fix $\{\mu^k\}$ and $\{\sigma^k\}$ at the values $\vartheta_3^{(\ell-1)}$ and $\vartheta_4^{(\ell-1)}$, respectively, and then find values of $(N_0, \{a_k\})$ that minimize the value of the cost functional J_2 using $(\vartheta_1^{(\ell-1)}, \vartheta_2^{(\ell-1)})$ as an initial iterate. Store the optimal value of N_0 in $\vartheta_1^{(\ell)}$ and the optimal value of $\{a_k\}$ in $\vartheta_2^{(\ell)}$.

7. Fix $\{a_k\}$ at the value $\vartheta_2^{(\ell)}$ and then find values of $(N_0, \{\mu^k\}, \{\sigma^k\})$ that minimize the value of the cost functional J_2 using $(\vartheta_1^{(\ell)}, \vartheta_3^{(\ell-1)}, \vartheta_4^{(\ell-1)})$ as an initial iterate. Store the optimal value of N_0 in $\vartheta_1^{(\ell)}$ (replacing the value from Step 6), the optimal value of $\{\mu^k\}$ in $\vartheta_3^{(\ell)}$, and the optimal value of $\{\sigma^k\}$ in $\vartheta_4^{(\ell)}$.

8. Increment ℓ by 1.

9. If $\ell < 20$, return to Step 6. Otherwise, terminate the algorithm.

B.4.1 Symmetric Cell Division

In the case of symmetric cell division, we can use (2.4) and (B.4) to write

$$\begin{aligned} n(t,x) &= \sum_{i=0}^{i_{\max}} N_i(t)\bar{n}_i(t,x) \\ &= \sum_{i=0}^{i_{\max}} N_i(t) \sum_{k=1}^{k_{\max}} a_k \text{logn}(x;\mu_i^k(t),(\sigma^k)^2). \end{aligned}$$

This quantity gives the number of cells per unit FI at time t with a *CFSE-induced* FI of x. Following the discussion in Section 2.2, we can use a convolution integral to obtain the structured density (in cells per unit FI) with an *observed* FI of $\tilde{x}$ (which incorporates FI due to autofluorescence). That is, using (2.10) we can write

$$\begin{aligned} \tilde{n}(t,\tilde{x}) &= \int_0^{\tilde{x}} n(t,x) f_{X_a}(\tilde{x}-x;t)\,dx \\ &= \int_0^{\tilde{x}} \left[\sum_{i=0}^{i_{\max}} N_i(t) \sum_{k=1}^{k_{\max}} a_k \text{logn}(x;\mu_i^k(t),(\sigma^k)^2)\right] f_{X_a}(\tilde{x}-x;t)\,dx \\ &= \sum_{i=0}^{i_{\max}} N_i(t) \sum_{k=1}^{k_{\max}} a_k \int_0^{\tilde{x}} \text{logn}(x;\mu_i^k(t),(\sigma^k)^2) f_{X_a}(\tilde{x}-x;t)\,dx. \end{aligned} \tag{B.9}$$

By assumption (cf. Section 2.5), $f_{X_a}(\xi;t)$ is a lognormal pdf in the argument ξ. Thus, each of the integrals in the last line of (B.9) is the convolution formula [18] giving the pdf of a random variable that is the sum of two lognormally distributed random variables. As proposed by Fenton [25], such a pdf can be approximated by a lognormal distribution with the same mean and variance as the actual distribution of the sum. (Note that this is discussed at length in Chapter 4.) Therefore, following the work of Hasenauer et al. [27], we use Fenton convolution approximations to arrive at

$$\tilde{n}(t,\tilde{x}) \approx \sum_{i=0}^{i_{\max}} N_i(t) \sum_{k=1}^{k_{\max}} a_k \text{logn}(x;\hat{\mu}_i^k(t),(\hat{\sigma}_i^k(t))^2),$$

where

$$\hat{\mu}_i^k(t) = \log(E_i^k(t)) - \frac{1}{2}\log\left(1+\left(\frac{SD_i^k(t)}{E_i^k(t)}\right)^2\right)$$

and

$$\hat{\sigma}_i^k(t) = \sqrt{\log\left(1+\left(\frac{SD_i^k(t)}{E_i^k(t)}\right)^2\right)}$$

are the parameters defining each of the $i \cdot k$ lognormal distributions with means and standard deviations given by

$$E_i^k(t) = \exp\left(\mu_i^k(t) + \frac{(\sigma^k)^2}{2}\right) + \mathrm{E}\left[X_a\right]$$

and

$$SD_i^k(t) = \sqrt{\left(\exp\left((\sigma^k)^2 - 1\right) \cdot \exp\left(2\mu_i^k(t) + (\sigma^k)^2\right)\right) + \left(\mathrm{SD}\left[X_a\right]\right)^2},$$

respectively.

B.4.2 Asymmetric Cell Division

As discussed in Section B.1.2, it is not, in general, possible to express $\bar{n}_i(t,x)$ as a linear combination of lognormal pdfs for arbitrary t when considering asymmetric cell division. Therefore, the convolution integrals involved in (2.10) cannot be distilled into convolutions of lognormal pdfs as in the symmetric division case. Instead, we use (2.4) to write

$$n(t,x) = \sum_{i=0}^{i_{\max}} N_i(t)\bar{n}_i(t,x),$$

where each of the pdfs $\bar{n}_i(t,x)$ can be evaluated recursively as described in Section B.1.2. Then, using (2.10) we can write

$$\begin{aligned}
\tilde{n}(t,\tilde{x}) &= \int_0^{\tilde{x}} n(t,x) f_{X_a}(\tilde{x}-x;t)\,dx \\
&= \int_0^{\tilde{x}} \left[\sum_{i=0}^{i_{\max}} N_i(t)\bar{n}_i(t,x)\right] f_{X_a}(\tilde{x}-x;t)\,dx \\
&= \sum_{i=0}^{i_{\max}} N_i(t) \int_0^{\tilde{x}} \bar{n}_i(t,x) f_{X_a}(\tilde{x}-x;t)\,dx. \qquad \text{(B.10)}
\end{aligned}$$

Now, each of the integrals in the last line of (B.10) is the convolution formula giving the pdf of a random variable that is the sum of two independent random variables (only one of which is lognormally distributed). Since we are not working with pairs of lognormal pdfs, we cannot use the Fenton approximations employed in the symmetric division case. We therefore use the trapezoid rule convolution method outlined in Algorithm 4.1.2 to evaluate

$$\int_0^{\tilde{x}} \bar{n}_i(t,x) f_{X_a}(\tilde{x}-x;t)\,dx$$

for each $i \in \{0,\ldots,i_{\max}\}$.

B.5 Details of Inverse Problem Implementation

For our purposes, we used the MATLAB routine `fmincon` (with the `active-set` optimization algorithm specified) to solve the minimization problems in steps 1 and 4 of Algorithm 2.6.1. Also, we set

$$\vec{q}_{typ} = (100, 100, 0.001, 1, 1, 1, 1, 1, 1, 0.1, 1, 1)$$

and chose $\varepsilon = 0.05$ for our implementation.