ABSTRACT

THOMPSON, WILLIAM CLAYTON. Partial Differential Equation Modeling of Flow Cytometry Data from CFSE-based Proliferation Assays. (Under the direction of H.T. Banks.)

The mammalian immune system is comprised of a complex network of cells which interact with each other as well as with external stimuli, and an immune response is characterized by the rapid proliferation (via division) of lymphocytes following exposure to some stimulating agent. Flow cytometric analysis of a proliferating cell population is a powerful and popular tool for the study of cell division and division-linked changes in cell behavior, as it permits the quick assessment of the phenotypic properties of a culture of proliferating cells. In particular, the development of the intracellular dye carboxyfluorescein succinimidyl ester (CFSE) [82] for the fluorescent labeling of cells has led to the need for quantitative models of division dynamics.

Some key features of a mathematical description of an immune response are an estimate of the number of responding cells and the manner in which those cells divide, differentiate, and die. Numerous mathematical treatments of CFSE flow cytometry data have been proposed to describe an immune response [38, 40, 51, 57, 62, 71], and each is motivated by the desire to relate the estimated numbers of cells in the population to average rates of division and death. Alternatively, the investigations of [75, 77] contain a structured partial differential equation (PDE) model (with CFSE fluorescence intensity as the structure variable) which can be fit directly to flow cytometry data.

After reviewing the data collection process and describing previous mathematical work, we focus on the application of such structured PDE models to CFSE histogram data. Several extensions and modifications of previous models are discussed and suggestions are presented to improve the agreement between model solutions and experimental data as well as to improve the physiological understanding of the model parameters. Next, the resulting structured PDE model is generalized into a system of PDE models representing the compartmentalization of the population of cells in terms of the number of divisions undergone since the beginning of the experiment. Mathematical aspects of this compartmental model are discussed, and the model is fit to a data set. It is shown that the compartmental model permits the quantification of cell counts in terms of the number of divisions undergone, so that key biological parameters such as population doubling time and precursor viability can be determined.

Finally, statistical models for the observed variability/noise in CFSE histogram data are discussed with implications for uncertainty quantification. It is revealed that several commonly held assumptions regarding the data collection procedure are not accurately reflected in the actual data. Using several additional data sets, experimental, intra-individual, and inter-individual variability in CFSE histogram data is qualitatively analyzed. The data collection procedure is then reexamined and a new statistical model of the data is hypothesized.

The models presented produce meaningful quantitative descriptions of the behavior of a dynamic population of cells and are sufficiently general to describe a wide array of proliferative behavior. Several generalizations of these models are also discussed with an eye toward experimental application. This work constitutes a significant first step toward the meaningful analysis of an immune response, and could provide a useful complement in experimental or diagnostic studies of the immune system.

Partial Differential Equation Modeling of Flow Cytometry Data from CFSE-based Proliferation Assays

by
William Clayton Thompson

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

2012

APPROVED BY:

Robert Martin	Ralph Smith
Hien Tran	H.T. Banks Chair of Advisory Committee

DEDICATION

To my wife, for her unwavering love, patience, and support through my graduate (and undergraduate) experience. And to my family, for their love and encouragement.

BIOGRAPHY

The author began his life in the small town of Ahoskie, North Carolina, where he spent nearly all of his time organizing pick-up games of baseball, football, basketball, and soccer with his younger brother and neighborhood friends. After more than ten years in Ahoskie, he and his family moved down the road to the tiny community of Harrellsville, to a house on the Chowan River. There, his summers were spent swimming, fishing, boating, and (eventually) skiing. Once at Hertford County High School, the author played soccer and baseball, eventually kicking for the football team his senior year.

While school was far from his favorite activity, it seemed to be a logical career choice given the author's prospects in sports or music. In the fall of 2004, the author became a Tarheel at the University of North Carolina in Chapel Hill. After two years of study and some soul-searching, the author dumped his physics major in favor of mathematics. Two years after that, he received a B.S. in mathematics with a minor concentration in physics. Graduate school seemed like an obvious next step. After a Recruitment Weekend visit to North Carolina State University, the author was so enamored that he decided to join his high school sweetheart (whom he had previously failed to convince to forsake her Wolfpack upbringing and join him at UNC) in Raleigh. The decision turned out well, as the author married his sweetheart two years later, and received his doctoral degree one year after that.

ACKNOWLEDGEMENTS

The work presented here is the culmination of 26 years (and counting!) of learning experiences in which far too many people played a part for their names to be listed individually. I would like to acknowledge everyone who has been instrumental in shaping my life to this point: my wife, whose love and support I rely on daily; my family, who have always provided for and encouraged my academic pursuits; my friends, who often provide a much-needed break from the office; and my teachers, who each share some role in making this research possible. Of course, a special acknowledgement goes to my Ph.D. advisor, who more than anyone else is directly responsible (some might say culpable) for guiding my professional development.

This research was supported in part by the National Institute of Allergy and Infectious Disease under grant NIAID 9R01AI071915, in part by the U.S. Air Force Office of Scientific Research under grant AFOSR-FA9550-09-1-0226. I am also grateful for additional support in the form a Center for Quantitative Science in Biomedice (CQSB) Fellowship and a Center for Research in Scientific Compution (CRSC)/Lord Corporation Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Mathematical Modeling of Flow Cytometry Data

1.1 Motivation

A quantitative understanding of cell division dynamics is of fundamental importance in numerous applications, from cancer and infectious disease diagnosis and treatment to immunosuppression therapies for transplant patients. These applications depend upon the accurate characterization of the rates at which cells divide, differentiate, and die and, of equal importance, how changing intra- and extracellular conditions affect these rates. In particular, the human immune response is a complex process in which the behavior of individual cells in the lymphatic system is altered by a multitude of intra- and extracellular signals. Thus, meaningful quantification of lymphocyte dynamics associated with clonal expansion during an immunoassay constitutes a significant step toward understanding the complex underlying processes of the biological system. The problem of studying these complex processes is two-fold. First, there is a need for an experimental procedure which can quickly and accurately provide division-related information in a population of dividing lymphocytes. Second, there is a need for mathematical models which can describe the data obtained from such a procedure. Thus the quantitative analysis of cell division is an important problem at the intersection of biology and mathematics.

Unlike many other cell types, the inherent mobility in vivo and nonadherence in vitro of lymphocytes makes the accurate determination of lineage very challenging [80] (although new techniques have been developed for this purpose [58]). In the absence of such data, one can look instead at the total number of divisions a cell has undergone since activation and how cells in different generations differ in phenotype (regardless of exact lineage). In the past two decades, a number of different techniques have been used for the study of cell growth and division [89, 104]. Early techniques, such as tritiated thymidine (^{3}H-Tdr) or bromodeoxyuridine (BrdU) uptake, while providing information regarding the fraction of dividing cells, are dependent upon cellular activation and do not provide information regarding how many divisions cells have undergone [82]. Other techniques, such as dimethylthiaxol (MTT), are sensitive to the activation state of the cells being labeled [28]. Experiments have also considered DNA content [24],

telomere length, and T-cell Receptor Excision Circles [30]. Lipophilic dyes which are incorporated into cellular membranes, such as PHK26, have been used successfully for the study of cell division history, although the uneven partitioning of the dye during mitosis can result in subsequent generations which are hard to distinguish [89, 104].

Since it was first described in 1994 [82], serial dilution of the intracellular fluorescent dye carboxyfluorescein succinimidyl ester (CFSE) for use in proliferation assays has become an essential tool for the determination of cellular division histories. CFSE is introduced into a culture of cells as carboxyfluorescein diacetate succinimidyl ester (CFDA-SE) which freely diffuses across the cell membrane and inside the cells. The acetate groups are then removed by intracellular esterases, producing highly fluorescent, membrane impermeant CFSE [89, 92]. CFSE is nonradioactive and stably incorporated (so that measurable concentrations remain within a viable cell for several weeks in vivo); it provides quick, bright, and approximately uniform labeling of all cells in a population (regardless of cell type or activation) [82, 104] and generally does not adversely affect the functioning of the intracellular machinery [78, 81, 82]. With a peak absorption at 491nm and a peak emission at 517nm, CFSE is compatible with standard fluorescein cytometry setups [89, 92]. Using a flow cytometer, the fluorescence intensity (a surrogate for CFSE content) of individual cells can be measured quickly and efficiently. Because the CFSE content of a cell is divided approximately in half at mitosis, the number of divisions a cell has undergone can be determined by comparing the measured fluorescence intensity of a cell to the measured fluorescence intensity of an undivided cell [80, 82, 89, 92, 104, 107]. When individual cell fluorescence intensity measurements for all cells in a given population are binned into a histogram, each generation of cells appears as a "peak" in the histogram data.

The compatibility of CFSE with other dyes renders possible the simultaneous measurement of division history and many other quantities, such as surface marker expression, cytokine content, and gene expression [80, 82, 104]. Most commonly, the proliferative characteristics of a cell population are measured in terms of the number of cells having undergone a specified number of divisions, plus any division-linked changes which are observed. The research presented here focuses primarily on the determination of cell proliferation and death rates (in terms of the number of divisions undergone) from flow cytometry histogram data for a population of lymphocytes, postponing any considerations of cell differentiation or division-linked changes. It should be emphasized, however, that the mathematical and experimental frameworks presented here apply more generally. In fact, the simultaneous measurement of multiple division-dependent properties combined with the applicability of the experimental technique to a wide variety of cell types has potentially profound applications in oncology (cancer metastasis and differentiation from normal cells), virology (latent viruses, HIV), and immunology (allergens, tissue grafting), either in the context of an interpretive framework, as a diagnostic tool, or even as part of a control mechanism (see, e.g. [27, 54, 57, 66, 67]).

Such uses for CFSE-based proliferation assays are premised upon an underlying model which can be used to establish a meaningful quantitative comparison of data sets in different experimental and biological conditions. Given the many desirable features of CFSE for cellular labeling and its near ubiquity in proliferation assay experiments, the research presented here focuses exclusively on flow cytometry data in which the cells have been labeled with CFSE–no distinction will be made between 'flow cytometry data' and 'flow cytometry data from CFSE labeling'. It should be noted that the mathematical analysis

presented here is sufficiently general to apply to any intracellular (or even lipophilic) fluorescent label for which the fluorochromes are divided approximately evenly upon mitosis. A survey of alternative techniques can be found in [89, 104]. Some additional techniques have also been considered in the introduction of [28].

The research presented here also focuses exclusively on data sets collected with CD4+ and CD8+ lymphocytes, as these are two important subsets of cells involved in an immune response. Yet it should be emphasized that the mathematical models considered are not specific to these cell types. Provided the cells are uniformly labeled with CFSE (which depends upon the distribution of certain types of intracellular proteins within the cells [84]) and that the intracellular proteins to which CFSE binds are partitioned approximately evenly upon mitosis, the models presented should apply.

In the remainder of this chapter, the mathematically relevant aspects of the data collection procedure are summarized and a data set is presented which will be the basis for the subsequent mathematical modeling efforts. Previous mathematical work is then reviewed. In Chapter 2, a structured partial differential equation model is presented which can be used to accurately fit CFSE flow cytometry data. Such a model is original in that is applied directly for CFSE histogram data (as opposed to numbers of cells having undergone a specified number of divisions, which must be approximated from histogram data by some analysis of the data), is physiologically motivated, and fits the data with unprecedented accuracy. This model is then generalized in Chapter 3 and biologically meaningful descriptors of the cell population are presented. Finally, statistical properties of flow cytometry histogram data sets are examined in Chapter 4. Generalizations of the model, applications, and future work are surveyed in Chapter 5.

1.2 Flow Cytometry Data from CFSE-Based Assays

Numerous protocols for the application of CFSE-based proliferation assays are available, and these protocols can be tailored to the specific goals of the experimenter [80, 82, 92, 107]. An original data set from a study of CD4+ lymphocytes is shown in Figure 1.1. This data set is the result of an in vitro proliferation assay with human blood mononuclear cells (PBMCs) isolated from a healthy blood donor. Approximately 5×10^6 to 5×10^7 cells were stained with $5\mu M$ CFDA-SE which is membrane permeable and taken into the cells by free diffusion. After initial labeling, the cell culture was washed to eliminate any excess CFSE. It is assumed that unbound CFSE in culture during the experiment, either left over from the initial labeling or resulting from apoptotic cells, is negligible. The population of cells was then stimulated to divide with a saturating quantity (in this experiment, 2.5 $\mu g/mL$) of phytohaemagglutinin (PHA). The time at which PHA is introduced to the cell culture is considered $t = 0$ hours. The cells were plated in 24 well plates at 1×10^6 cells/mL RPMI-1640/10% FCS nutrient medium. Beginning at day 3, every 24 hours one third of the medium was exchanged with fresh medium to ensure sustained cell nutrition.

At each sample time, cells from a single well are harvested and transferred to Trucount tubes containing 51466 calibration beads. (The use of separate wells prevents the disruption of the proliferating cell populations. It is tacitly assumed that each well plate contains an identical population of cells at all times.) Cells were then stained with fluorescently labeled anti-CD4 antibodies. This staining makes

it possible to distinguish the CD4+ cells from other cells in the PBMC culture and does not disrupt the measurement of fluorescence intensity resulting from intracellular CFSE. These cells are then analyzed by a flow cytometer. A flow cytometer uses laser light to measure various properties of an individual cell. As cells are forced one at a time through the measurement apparatus, attributes such as size, granularity, CFSE content, and surface marker expression are recorded. A flow cytometer is capable of measuring thousands of cells in a single second, so any changes to the measured population during the measurement process are negligible.

Because of physical limitations, a flow cytometer will measure only a fraction of the total contents of a particular well. For this reason the Trucount tubes contain the known number of calibration beads which can be easily detected in the flow cytometry output. By counting the number of beads measured by the flow cytometer and comparing this number to the total number of beads originally in the culture, one can scale up the number of counted cells (in the flow cytometer output) to obtain an estimate of the total number of cells in the well plate at the measurement time. Thus we assume that the sample of cells analyzed by the flow cytometer is representative of the population, and that the scaling accurately reflects the actual number of cells. It should also be noted that the PBMC culture measured by a flow cytometer contains a large number of cells (e.g., B-cells, monocytes, CD8+ T-cells) which are not of interest for the current study. These cells can be 'gated out' by the experimenter based upon known properties (size, granularity, surface marker expression, etc.) so that they are excluded from analysis. It is assumed that our data set contains only CD4+ cells, and that all CD4+ cells in the population are represented in the data. While this report focuses exclusively on human CD4+ lymphocytes cultured in vitro, CFSE-based assays have been used successfully in mice (both in vivo and in vitro) and on a wide variety of cell types, including other T lymphocytes, B lymphocytes, NK cells, bacteria, fibroblasts, hematopoietic stem cells, and smooth muscle cells [80, 89, 92, 105]. More information regarding the CFSE protocol in general can be found in [80, 82, 92, 104, 107].

Qualitatively, the flow cytometer returns a measure of the fluorescence intensity (FI) of a given cell, owing primarily to the presence of CFSE within the cell. In order to obtain this measurement, the flow cytometer uses hydrodynamic focusing to push cells one at a time through a beam of laser light. This light is absorbed and then emitted again by the CFSE molecules within a cell. This emitted light is filtered and then quantified by a photometer. It is known that measured CFSE FI has a varies approximately linearly with the concentration of CFSE used in the staining process [80, Fig. 3]. Because measured CFSE FI does not change as cells become activated and swell [80, Fig. 6] it is expected that CFSE FI is correlated with the mass of CFSE within a cell. While most of the measured FI of a given cell is the result of CFSE, all cells have a small but measurable autofluorescence intensity (AutoFI) as a result of the spectral properties of naturally occuring intracellular molecules. As will be shown in Chapter 3, AutoFI is a subtle but important issue which must be addressed in a mathematical model.

While measurements have been made for each individual cell, the population of cells is most commonly represented in a series of histograms which present the number of cells analyzed by the flow cytometer having measured CFSE FI in a given range. The data is stored as a set of ordered pairs (z_k^j, n_k^j), $k = 1, \ldots, K(j)$ which corresponds to the number of cells n_k^j counted into the bin with left boundary z_k^j at time t_j. The notation is meant to emphasize the possibility that the histogram bins need not share a common fixed width, nor need they be the same at each measurement time. It is this histogram data

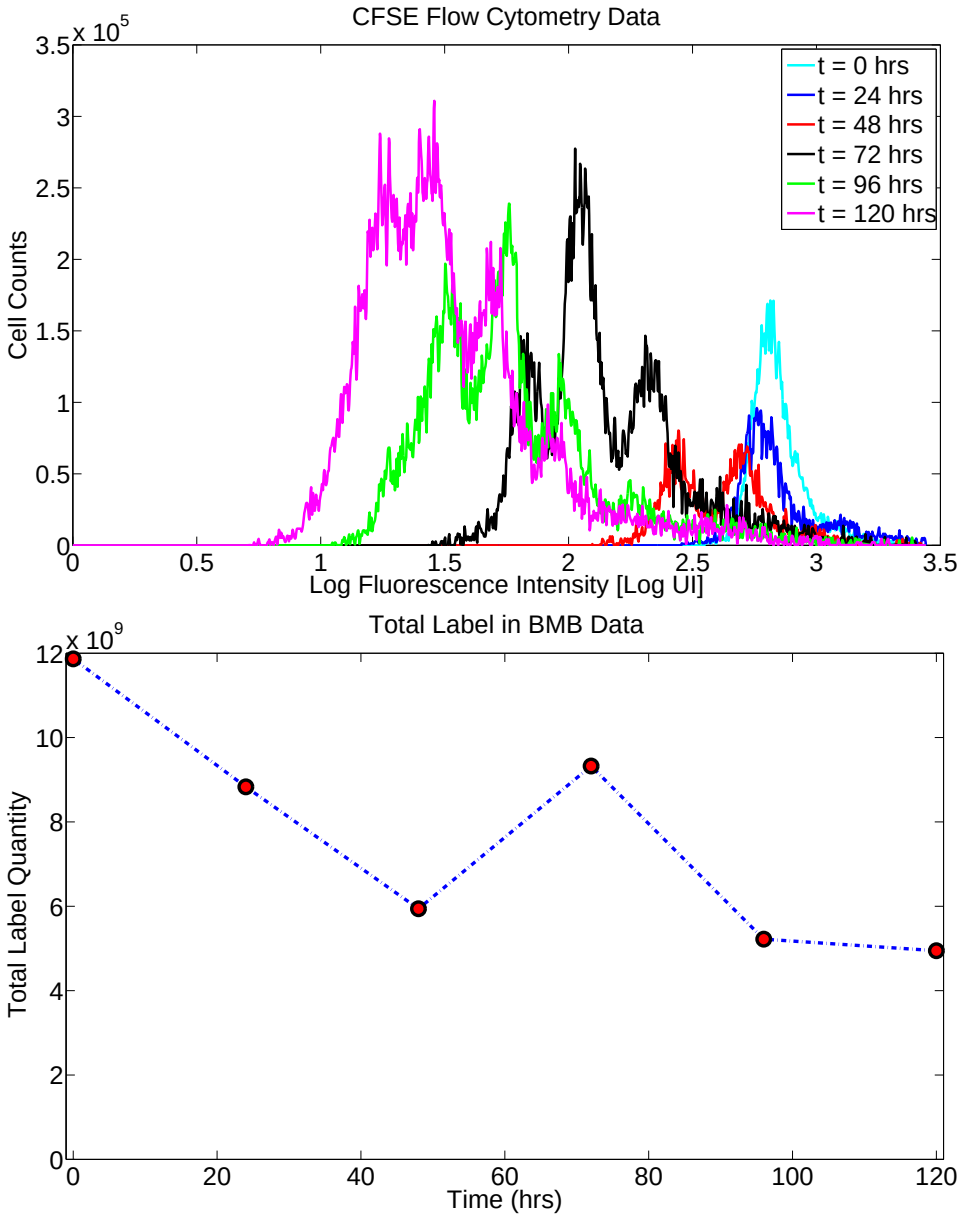

Figure 1.1: Top: A collection of histograms for CFSE flow cytometry data from [20, 21, 77]. Each histogram corresponds to a collection of cells measured at the same time. Thus the data (for which a mathematical model is sought) contains the numbers of cells counted into each histogram bin at each measurement time. Bottom: Total computed label content $\int zn(t,z)dz$ over time for the data from [21]. The increase at $t = 72$ hours violates commonly held assumptions regarding the data collection procedure.

for which we seek a mathematical model. In general, it is possible to choose the bins for the histograms as one wishes. But the current data set (graphically displayed in Figure 1.1) was already reported as histogram data in [77] when we began efforts on it [21]. While this may have implications for a statistical error model (see Section 4.5), it is of little consequence for the moment.

Because CFSE FI divides approximately in half with each subsequent division (at least for the first few generations–see Section 2.1) it is most convenient to use a logarithmic scale for CFSE FI. The initial uptake of CFSE is relatively uniform across the population of cells being studied so that the initial distribution of cells appears as a unimodal 'peak' on a histogram. When cells divide, the mass of CFSE within each daughter cell is approximately half that of the original parent cell. The corresponding decrease in measured CFSE FI results in the emergence of a second 'peak' in the data. As cells continue to divide, the initially unimodal histogram data becomes multimodal, with each mode representing a distinct generation of cells. Over time, all cells (even in the absence of division) slowly drift to the left, reflecting a loss of label. An effective mathematical model must adequately describe both the emergence of these distinct peaks as well as the slow decay of the label.

While the results in [21, 77] were generated by fitting to the data from all five measurement times ($t = 24$, 48, 72, 96, and 120 hours), the current study will not make use of the data at $t = 72$ hours. Because CFSE is added to the cell culture at the beginning of the experiment but not afterward, the total mass of CFSE FI in culture cannot increase over the course of the experiment. (While the separate measurements in time are obtained from distinct samples of cells in separate wells, the assumption that the histograms sufficiently represent a single population is standard.) Because fluorescence intensity is approximately proportional to the mass of CFSE within a cell, the sum of all cells in a population, weighted by measured FI, provides an indication of the mass of CFSE within the measured population. We have found that this 'total label content' is greater at $t = 72$ hours than at the previous time point (see Figure 1.1), indicating a net increase in the mass of CFSE between $t = 48$ hours and $t = 72$ hours. It should be emphasized that such an anomaly in CFSE data sets is not the result of any correctable errors in the experimental process and has been noticed elsewhere [38, 62]. As will be shown in Chapter 4, this feature is explained by a more complete statistical model of the data which accounts for the manner in which a sample of cells is measured as a surrogate for the entire population of cells. At the moment, it is assumed that this anomaly is the result of measurement or scaling error or some unknown and unmodeled biological event and thus the $t = 72$ hours data point will not be included in the present investigation.

The data set we will use to calibrate the compartmental model is shown in Figure 1.2, with measurements taken at $t = 24$, 48, 96, and 120 hours. (Data from $t = 0$ hours is used to form an initial condition for the mathematical model.) There is a small cohort of cells with high CFSE FI ($\log \text{FI} \geq 3$) visible in the data at $t = 24$ and 48 hours. These cohorts are believed to be cell duplets (cells which are clumped together) or otherwise anomalous cells which were not gated out of the population data. Such cells are are unmodeled, and their presence in the data will have some effect on the parameter estimation procedure to be discussed. However, the effect of these cells is small and it is not believed that similar features will be observed in additional data sets. Thus the failure of the model to fit these cells is generally ignored except where a comment is necessary.

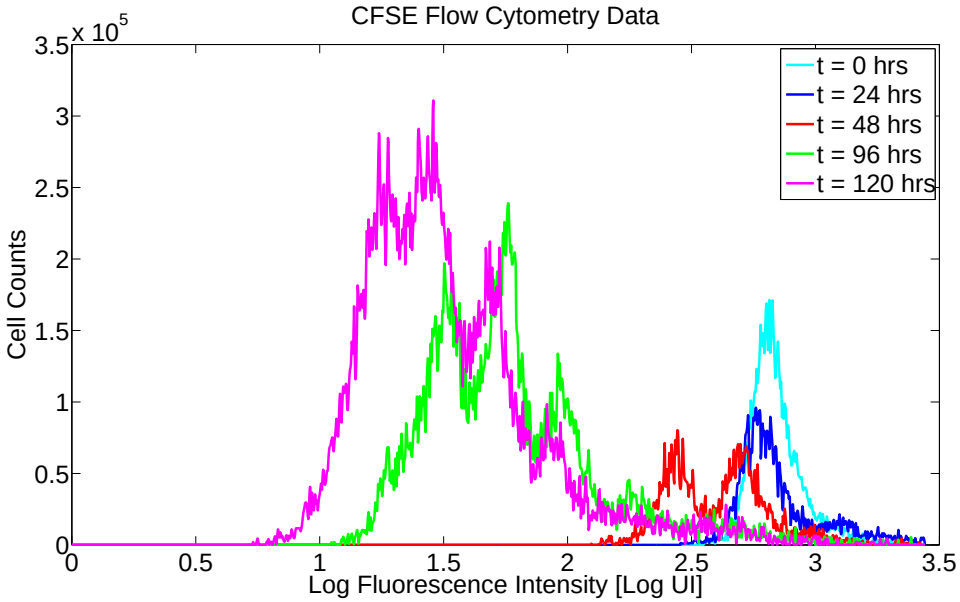

Figure 1.2: A collection of histograms for CFSE flow cytometry data. Each histogram corresponds to a collection of cells measured at the same time. Thus the data (for which a mathematical model is sought) contains the numbers of cells counted into each histogram bin at each measurement time. From [20, 21].

1.3 Survey of Mathematical Models of CFSE Data

Because the flow cytometer provides measurements of individual cells within a sample, the mathematical analysis of lymphocyte activation and division can be performed on a wide range of scales, from the molecular level (antigen presentation and recognition) to the population level. However, because individual cells (or their progeny) are not followed over the course of the experiment, the measurement process is in effect an aggregate sampling, and population level modeling seems more appropriate. The primary motivation behind population level modeling is that, while the behavior of individual cells may be highly variable as a result of sensitivity to any number of random intracellular and/or environmental factors, the immune response as a whole (to be understood as the aggregate behavior of all cells in the population) is regular and predictable [99].

Early experimental techniques, such as ^{3}H-Tdr or BrdU labeling, were limited to the consideration of the proportion of dividing cells in a population. From such data, it is possible to obtain basic information regarding the proliferative capacity of a sample of cells. However, such descriptive and semi-quantitative methods are generally restricted to populations of cells which divide synchronously [104, 107]. Yet asynchronous division is a well-known feature of an immune response [51]. A more quantitative analysis is possible if the peaks in the histogram data (see Figure 1.1) are analyzed in some way to determine the numbers of cells in each generation. This can be done most simply by interval gating (see, e.g., [87]), where the FI or log FI axis is partitioned into intervals which approximately correspond to distinct generations of cells. More accurately, the CFSE data can be fit (typically in a least-squares sense) with

gaussian or log-normal curves, with the area under each curve used to determine the numbers of cells in each generation [80, 92].

Mathematical models of CFSE data have focused almost exclusively on fitting such numbers in an attempt to examine how various experimental conditions alter an immune response. The literature on the mathematical modeling of CFSE data falls into two broad classes. First, there are 'cellular calculus' models [51] which aim to describe the net effects of experimental changes on the observed population behavior without regard for the exact cellular and/or molecular mechanisms responsible for those changes. The second class of models attempt to motivate behavioral changes as direct consequences of mechanisms within the cell cycle. It can be shown that these two classes of models are derivable from different assumptions regarding the nature of variability in the population of cells (although neither class of model is strictly dependent upon those assumptions). The first class of models can be derived from the assumption that variability is inherent in the individual cells themselves, while the second class can be derived from the assumption that variability arises from stochastic environmental interactions for otherwise identical cells [73].

One of the earliest cellular calculus models for asynchronously dividing cell populations (from which the phrase 'cellular calculus' is taken) is the work of Gett and Hodgkin [51]. Using cell numbers determined by fitting CFSE histogram data with a series of log-normal curves, they computed the total numbers of cells in each generation over the course of several days and tracked how these numbers changed over time. It was shown that, after one division, normalized precursor frequency (see below) as a function of the number of divisions undergone could be well-fit by a gaussian curve. Thus it was hypothesized that the primary source of asynchrony within the population was a normally distributed time-to-first-division. The mean of the gaussian curve, which reflected the average number of divisions undergone for proliferating cells, was seen to increase linearly with time and the rate of increase provided a measure of the average proliferation rate. The authors go on to compute parameters such as mean division rate and mean time to first division, and to see how these parameters change in various stimulation conditions.

While the method of Gett and Hodgkin is nonintuitive, it can be shown [39] that the parameters estimated follow directly from an ordinary differential equation (ODE) model of cell division. Let $N_i(t)$ be the total number of cells having completed i divisions at time t. Let α be the (exponential) rate at which cells divide, and let β be the (exponential) rate at which cells die. R.J. de Boer et al. begin by considering the ODE model of Revy et al.,

$$\begin{aligned}\frac{dN_0}{dt} &= -(\alpha+\beta)N_0(t)\\ \frac{dn_i}{dt} &= -(\alpha+\beta)N_i + 2\alpha N_{i-1}(t).\end{aligned}$$

It follows that the solution for each generation of cells is

$$\begin{aligned}N_0(t) &= N_0(0)e^{-(\alpha+\beta)t}\\ N_i(t) &= \frac{(2\alpha t)^i}{i!}N_0(t).\end{aligned}$$

A precursor is a cell originally (at $t = 0$) in the population under study which gives rise to some additional number of cells via mitosis. Let $P_i(t)$ be the total number of precursors for the cells in generation i at time t. It follows that $P_i(t) = N_i(t)/2^i$ (allowing for fractional precursors). Thus we have

$$\begin{aligned} P_0(t) =& N_0(t) \\ P_i(t) =& \frac{(\alpha t)^i}{i!} N_0(t). \end{aligned}$$

The total number of precursors at time t is

$$P(t) = \sum_i P_i(t) = N_0(0) e^{-(\alpha+\beta)t} \left(\sum_i \frac{(\alpha t)^i}{i!} \right) = N_0(0) e^{-\beta t}.$$

Finally, the *normalized precursor frequency* for generation i is

$$\frac{P_i(t)}{P(t)} = \frac{(\alpha t)^i}{i!} e^{-\alpha t},$$

which is a Poisson process with mean αt [39]. The method of Gett and Hodgkin [51] uses a plot of the Poisson process $P_i(t)/P(t)$ versus i, which will appear normal for sufficiently large t [38]; as noted above, the mean of this distribution increases linearly with rate α so that the rate of proliferation can be easily determined. The Gett and Hodgkin model has been generalized by de Boer et al. to include an initial transient as well as death rates which change with division number [39].

A major strength of the Gett and Hodgkin model is the manner in which intuitive, biologically meaningful parameters can be determined almost directly from graphical representations of the data and with very little sophisticated mathematics. However, since its initial publication, several shortcomings of the model have been addressed. As shown above, the Gett and Hodgkin model is consistent with a simple ODE model for cell division. ODE models can be interpreted as 'random birth-death' models [50] in which all cells have a fixed probability of dividing or dying which does not depend upon the division history of the cells (specifically, on the time since last division). This amounts to a tacit assumption of a time to divide or die which is exponentially distributed without any imposition of a minimum cell cycle time [39]. ODE models have been used successfully in some situations [2, 93, 103], and may be particularly useful (given their simplicity) for data in which the generation peaks of the histogram data are poorly resolved [2]. But in general, the exponential distribution of division times can cause too much asynchrony in the computed distribution of cells among generations so that the mean time to first division cannot be accurately estimated [39, 50, 91]. Several analyses have found that a minimum cell cycle time is an essential feature for the accurate analysis of CFSE data [39, 50, 91], as models lacking this feature are generally limited to describing the mean number of divisions undergone [38, 39] for cells which divide slowly [50]. Experimental data typically exhibits a delay between activation and the onset of proliferation which cannot be modeled in an ODE framework [34]. A more detailed treatment of the assumptions, uses, and limitations of ODE models can be found in [34].

A revision of the Gett and Hodgkin model by Deenick et al. [40] addresses several features observed in flow cytometry data sets which could not be explained previously. In the revised model, a progressor

fraction is used to account for the subset of the initial population of cells which will not divide. A death rate for divided cells is incorporated into the estimation procedure and a fixed cell cycle duration Δ is also imposed. As in the Gett and Hodgkin model, it is assumed that the only source of asynchrony in the proliferating population is the time taken to enter the first division. Then the Deenick et al. model can be written [38] as an ODE-algebraic equation system,

$$\begin{aligned}\frac{dN_1(t)}{dt} =& R(t) - R(t-\Delta)e^{-\beta\Delta} - \beta N_1(t)\\ N_i(t) =& \left(2e^{-\beta\Delta}\right)^{i-1} N_1(t-(i-1)\Delta),\end{aligned}$$

where $R(t)$ is the density of cells which divide at time t and β is defined as before. The undivided cells are described by

$$N_0(t) = N_0(0) - \int_0^t R(t)dt.$$

A derivation can be found in [38].

Both the Gett and Hodgkin model and the Deenick et al. model are intuitive and establish parameters which are readily identifiable from data sets. Using these models, changes in parameters resulting from changing experimental conditions can be used to describe in exact terms how stimulatory and costimulatory conditions directly affect proliferative capacity. However, several studies have highlighted difficulties in the interpretations of the parameters estimated from the underlying models. For instance, a generalization of the Deenick et al. model has also been considered in which the rate of death changes with each division number [38]. There, it was shown that the effects of increased cell cycle time or increased death with division number are essentially indistinguishable. It can also be shown that the rate of linear increase of the mean division number (and hence, the estimated average cell cycle time) in the Gett and Hodgkin model depends upon the mathematization of cell death within the cell cycle, so that the procedure for determining biologically meaningful quantities is not robust to a model misspecification of cell death within the cell cycle [39, 50, 91].

In the past decade, researchers have shifted away from simple 'cellular calculus' models and toward mathematical models which more accurately reflect the biological mechanisms responsible for cell division. In many cases, these mathematical formalisms of cell cycle dynamics begin with the Smith-Martin [98] model of the cell cycle, which was originally proposed to account for an observed variability in cell cycle times. In the Smith-Martin model, the cell cycle is divided into a stochastic A state and a deterministic B state (corresponding approximately to the G1 and S-G2-M phases of the cell cycle, respectively). Cells may remain in the A state for an indefinite period of time, but exit with some fixed probability and enter the B state. The B state has a fixed duration, after which cells divide and return to the A state [98].

An early mathematization of the Smith-Martin model for application to CFSE-based flow cytometry data was considered by Nordon et al [87]. A model is presented in terms of precursor numbers (without regard for the expansion of total cell number with each division) and a set of biologically meaningful parameters are proposed which can be easily determined from histogram data (cf. Gett and Hodgkin [51]). A complete Smith-Martin model (accounting for total cell numbers rather than just precursor numbers) can be derived from a coupled set of equations [28, 91]. As in traditional ODE models, the

constant rate of transition from the A state to the B state is modeled with a linear ordinary differential equation and reflects an exponential distribution for the transition time. Meanwhile, the fixed duration of the B state can be modeled with an age-structured partial differential equation. Let $A_i(t)$ represent the total number of cells having completed i divisions at time t and currently in the 'A' state (approximately the G1 phase of the cell cycle). Let $b_i(t,s)$ be the age-structured density (number per unit age) of cells having completed i divisions at time t and having spent time s in the 'B' state. Then the coupled system is

$$\begin{aligned}\frac{dA_i(t)}{dt} =& 2b_{i-1}(t,\Delta) - (\alpha - \beta_A)A_i(t)\\ \frac{\partial b_i}{\partial t} + \frac{\partial b_i}{\partial s} =& -\beta_B b_i(t,s) \quad 0 \le s \le \Delta\\ b_i(t,0) =& \alpha A_n(t),\end{aligned}$$

where Δ is the (fixed) duration of the B phase of the cycle, α is the rate of transition from the A state to the B state, and β_A and β_B are the (constant) probabilities of death in the two states. By convention, $b_{-1}(t,s) = 0$ for all t, s. Trivially, we have the solution

$$b_i(t,s) = \alpha e^{-\beta_B s} A_i(t-s), \quad i \ge 0,$$

and thus the total number of cells in the A state can be solved inductively on the number of divisions undergone,

$$\frac{dA_i(t)}{dt} = -(\alpha + \beta_A)A_i(t) + 2\alpha e^{-\beta_B \Delta} A_{i-1}(t-\Delta), \quad i \ge 0.$$

(Again, it is assumed $A_{i-1}(t) = 0$ for all t.) Then the total number of cells having undergone i divisions is $N_i(t) = A_i(t) + B_i(t)$ where

$$B_i(t) = \int_0^\Delta b_i(t,s)ds = \alpha \int_0^\Delta e^{-\beta_B s} A_i(t-s)ds.$$

Alternatively, the cells in the B state can be computed [49] as

$$\frac{dB_i(t)}{dt} = \alpha A_i(t) - \alpha e^{-\beta_B \Delta} A_i(t-\Delta) - \beta_B B_i(t).$$

Interestingly, it can be shown that the ODE model of Revy et al [93] (acknowledged above as consistent with the model of Gett and Hodgkin [51]) is obtained in the limit as $\Delta \to 0$, provided $\beta_A = \beta_B = \beta$. However, under the alternative assumption that $\beta_A = 0$ while $\beta_B \neq 0$, a different ODE model is obtained and the parameters estimated by the graphical method of [51] will not accurately reflect the actual underlying dynamics [91]. The Deenick et al. model is equivalent to a Smith Martin model with no A state. The age-structured PDE then has the new boundary condition $b_i(t,s) = 2b_{i-1}(t,\Delta)$. An alternative to differential equations formulations of the Smith-Martin model has also been considered using agent-based modeling [36].

The Smith-Martin model has been shown to much more accurately describe cell counts obtained from CFSE data when compared to ODE models. This is most likely the result of the minimum cell

cycle time created by the delay in the Smith-Martin model [39]. In a simulation study, de Boer et al. show that many of the observed features which prompted the creation of the Gett-Hodgkin model [51] can be reproduced with a Smith-Martin model. Based upon the observation that the first division after activation typically takes much longer than subsequent divisions, the authors go on the propose a generalization of the Smith-Martin model to allow for heterogeneous (with respect to the number of divisions undergone) transition rates and cell cycle times. The resulting model is

$$\begin{aligned}
\frac{dA_0(t)}{dt} &= -(\alpha_0 + \beta_0)A_0(t) \\
\frac{dA_1(t)}{dt} &= -(\alpha + \beta_A)A_1(t) + 2\alpha_0 A_0(t - \Delta_0)e^{-\beta_B \Delta_0} \\
\frac{dA_i(t)}{dt} &= -(\alpha + \beta_A)A_i(t) + 2\alpha A_{i-1}(t - \Delta)e^{\beta_B \Delta}.
\end{aligned}$$

As before, the total number of cells in each generation is $N_i(t) = A_i(t) + B_i(t)$ where

$$\begin{aligned}
B_0(t) &= \alpha_0 \int_0^{\Delta_0} A_0(t - s)e^{-\beta_B s} ds \\
B_i(t) &= \alpha \int_0^{\Delta} A_i(t - s)e^{-\beta_B s} ds.
\end{aligned}$$

Such a heterogeneous model has been found to more accurately describe observed population dynamics when compared to a model in which transition rates and cycle times do not change after the first division [73]. As a further generalization, the exponential probability distribution for transition from the A state can be replaced with an arbitrary probability distribution [38]. For instance, Lee and Perelson test log-normal and gamma distributions for the probability of transition out of the A state, and find that a delayed gamma distribution can be derived from assumptions regarding a two-step activation process [72]. They go on to consider a 'generalized heterogeneous Smith-Martin model' which incorporates a progressor fraction to account for cells which do not divide, and cell cycle lengths which increase with division number [72]. The Smith-Martin model has also been generalized to allow for division-linked differentiation and the inheritance of division times [86].

Pilyugin et al. [91] draw parallels between the Smith-Martin model above and renewal equations used in demographic modeling. They then propose a number of invariant parameters and illustrate graphical techniques for estimating these parameters. Additional methods of parameter estimation for the Smith-Martin model are surveyed in [50]. Both direct (least squares) fitting of the model output to cell count data as well as indirect/graphical methods for parameter estimation suffer from a lack of identifiability in some of the parameters to be estimated. In particular, one cannot uniquely determine the rates of death, β_A and β_B in the two phases of the cell cycle without additional information or assumptions, particularly when the heterogeneous Smith-Martin model is used [50]. Similarly, Lee and Perelson find that the effects of a linear increase of cell cycle length with division number cannot be distinguished (using only cell count data) from the effects of an increasing death rate with division number [72].

Given the random nature of the mechanisms of cell division and death highlighted thus far, many recent models have placed an increasing emphasis on probabilistic structures to describe rates of division and death within a population. Using observations obtained from several experiments, Hawkins et al.

[57] showed that cell behavior is consistent with the hypothesis that the cellular controls of division and death operate independently of one another. It was hypothesized that the fate of a single cell could be described by two random variables: a time to divide and a time to die. With each division, daughter cells would be governed by a realization of the two parameters, with no inheritance of division or death times from the previous generation; the fate of each cell is determined by the smaller of the two random variable realizations. It follows that the expected number of cells having undergone a specified number of divisions at a given time can be determined from the probability distributions from which the times to divide and die and drawn, as well as the initial number of cells in the population. Assuming a log normal time to division and/or death, Hawkins et al. propose simple equations which accurately describe cell numbers obtained from CFSE data [57] and which outperform the classical Smith-Martin model [34, 57].

In simplest terms, the cyton model computes the number of cells having undergone i divisions at time t as

$$N_i(t) = 2^i N_0 \left(\int_0^t R_i(\tau)d\tau - \int_0^t D_i(\tau)d\tau - \int_0^t R_{i+1}(\tau)d\tau \right),$$

where $R(t)$ is the probability of recruitment into the i^{th} division at time t, and $D_i(t)$ is the probability of death for cells having undergone i divisions at time t. Thus, the total number of cells having undergone a specified number of divisions at a given time depends upon the form of the functions $\{R_i(t)\}$ and $\{D_i(t)\}$. Lee and Perelson show that, using the appropriate functions $\{R_i(t)\}$ and $\{D_i(t)\}$, the cyton model is equivalent to the generalized heterogeneous Smith-Martin model [71]. Thus the cyton model can be considered as a generalization of the existing Smith-Martin models.

In fact, the cyton model itself is a specific case of stochastic branching process models [62, 99]. Much like the cyton model, the behavior of cells in a branching process is described by the probabilities with which cells divide and die. An early discrete time branching process model for CFSE-based proliferation assays was considered by Yates et al. [109], with a particular interest in how a population of cells which arise in a branching process (so that cells with a common progenitor are mutually dependent) could be fit to data in a statistically rigorous manner. Hyrien and Zand [61] propose an age-dependent Bellman-Harris branching process and demonstrate the consistency of several parameter estimators, provided cells evolve independently of one another and cell death is negligible in the population. In a later paper, Hyrien et al. [62] propose that the cyton model amounts to a specific subset of branching process models they term 'competing risk' models. Using arguments from the theory of branching processes, they show that the cyton model tacitly assumes that a cell's 'decision' to divide or die is concurrent with the event itself. Meanwhile, general branching process models do not require such an assumption and were found to more accurately fit data for CD8+ cells. More recently, branching process arguments have been used to generalize Smith-Martin-type models to account for multiple cell types and the inheritance of dynamic parameters.

The determination of cell counts from such probabilistic models is effectively equivalent to considering the expectation of the probability structures. Higher order moments, such as the variance, can then be computed to provide some indication of the accuracy of parameter estimates obtained from using such models. Subramanian et al. use arguments from the theory of branching processes to compute variances for a generalization of the cyton model [99]. They find that, provided there is limited correlation between individual cells in the population, the variance of the cyton model around the mean is small and the

parameters of the cyton model can be reasonably estimated. Unfortunately, recent evidence suggests that correlation is nonneglgible between certain sets of cells [58]. Duffy et al. use branching process arguments to generalize the cyton model to include correlations between certain subsets of cells. In the simplest case, it is assumed that sibling cells share a common time to divide. The authors proceed to consider common division and death times for cousin and 2^{nd} cousin cells, and finally to common parameters for all cells in a given generation which share a common precursor. It has been shown that as the magnitude of the correlation between cells increases, the variance of the population around the mean also increases. The mean dynamics, however, are generally unchanged for sufficiently large populations of cells [62, 106].

1.4 Structured Population Models

Each of the models discussed thus far has been used effectively to provide various measures of the proliferative capacity of a population of cells. In general, these models are based upon estimation from the cell numbers computed by the deconvolution of CFSE histogram data with normal or log normal curves. Such approaches are straightforward and easy to implement, and the resulting cell numbers provide an accurate description of the distribution of cells in the population. However, the imposition of particular shapes for the generational structure of CFSE histogram data (during the deconvolution process) can introduce biased insight into the generation structure of the cells, and hence into the resulting division and death rates. Alternatively, we propose that there is information to be learned not only from modeling the total numbers of cells, but also from the direct modeling of the complete experimental process. This is a more fundamental level of analysis of the kinetics of cell turnover which we believe to provide a more accurate assessment of the biological processes occurring in the population. Given such a goal, the common use of histograms to represent CFSE flow cytometry proliferation assay data makes structured population models a natural framework in which to work. Significant literature exists on the subject of structured population models, going back at least as far as the Sinko-Streifer [97] model for general populations or the Bell-Anderson [26] model for volume-structured cell populations. More recently, "physiologically-structured" population models [85] have been developed for cell populations structured by age [1, 25, 43, 53], cyclin content [25], and size [44, 53, 54, 90] as well as DNA-content [24].

The measurement of CFSE FI by a flow cytometer makes measured fluorescence intensity a natural structure variable for a structured population model. While not a physiological variable, the notion that such a structure might be used to accurately model cytometry data by accounting for the natural dilution of label was proposed at least as early as the year 2000 for BrdU-based assays [29]. To our knowledge, Luzyanina, et al. [77] proposed the first model to explicitly employ fluorescence intensity as a structure variable in a partial differential equation (PDE) framework. There it was shown that such a model can be effectively used for the tracking of a proliferating lymphocyte population stained with CFSE, and that such a model is as effective as compartmental ODE models for estimating the numbers of cells having undergone a specified number of divisions. More recent work [7, 21, 75] has consistently demonstrated that the label-structured PDE framework can accurately model the observed histogram data from a CFSE-based proliferation assay. The key idea behind the use of FI as a structure variable is that, because CFSE FI decreases upon division, fluorescence intensity can be used as a surrogate for division number.

Thus the conclusions obtained with previous models [38, 57, 58, 71, 72, 73] regarding the necessity of division-dependent rates of proliferation and death can be assessed. We believe that the primary benefit of using such a model lies in its ability to treat the measured FI data directly, thus accounting for the intracellular dynamics of label dilution while simultaneously estimating proliferation and death dynamics at the population level. Moreover, this method relies less on distinct peak separations in the CFSE histogram data (see Section 3.6.1), a potential advantage when working with heterogeneous cell populations.

While these models are indeed effective, the parameter estimates which resulted from fitting these models to an available data set seemed to suggest that label was being created during the process of cell division, a known impossibility [21]. In the next chapter, we revisit the work presented in [21] and [77] and resolve two key problems addressed there. First, we explain the apparent creation of label during cell division. It is shown that this apparent physiological impossibility is actually readily explained and removed from the models by the inclusion of cellular autofluorescence. Second, by examining data from cells which were stained with CFSE but not stimulated to divide, we find that a minor modification of the exponential decay first proposed in [77] can provide a superior fit to the data. These two revisions (autofluorescence and biphasic label decay) provide important insights into the mathematical analysis of turnover kinetics for cells stained with CFSE and measured via flow cytometry. Their accurate modeling is vital to the meaningful estimation of population proliferation and death rates in a manner which is unbiased and mechanistically sound. Significantly, this new model is still sufficiently general to apply to a wide range of cell types and stimulation conditions.

In Chapter 3, this label structured model is revised further and used to compute the numbers of cells having proceeded through a fixed number of generations. It is shown that these cell numbers can in turn be used to calculate simple, biologically relevant parameters which can be used to quickly summarize an immune response for the experimental conditions under consideration. Thus, it should be possible to quantify the effects of extracellular conditions such as stimulation strength and duration on proliferative behavior (e.g., [40, 51]). Statistical aspects of the flow cytometry data are considered in Chapter 4 with implications for the quantification of uncertainty when model parameters are estimated from data. Generalizations, applications, and future modeling efforts are considered briefly in Chapter 5.

Chapter 2

A Label-Structured Partial Differential Equation Model

2.1 Summary of Initial Structured Model

We begin this chapter with an overview of the modeling results for the initial structured population model [21, 77]. A structured PDE model for the dynamics of the life and death process of a population of cells labeled with CFSE was initially proposed in [77] as a variation of a Bell-Anderson [26] or Sinko-Streifer [97] population model. Let x denote the CFSE FI (in units of intensity, UI) of a cell and let $n(t,x)$ be the label-structured population density (cells/UI) of cells with FI x at time t. Then the population density for $x \in [x_{\min}, x_{\max}]$ and $t > 0$ is governed by a hyperbolic partial differential equation

$$\begin{aligned} \frac{\partial n}{\partial t}(t,x) + \frac{\partial [v(x)n(t,x)]}{\partial x} = &-(\alpha(x)+\beta(x))n(t,x) \\ &+\chi_{[x_{\min}, x_{\max}/\gamma]} 2\gamma\alpha(\gamma x)n(t,\gamma x), \end{aligned} \tag{2.1}$$

where $\alpha(x)$ is the cell proliferation rate (hr^{-1}) and $\beta(x)$ is the cell death rate (hr^{-1}). The term $v(x)$ represents the natural label loss rate (UI/hr) as cells naturally lose FI over time even in the absence of division (due to catabolic activity [78]). The parameter γ is the label dilution factor, representing the ratio of FI of a mother cell to FI of a daughter cell. Thus the second term on the left represents the velocity of decay of the CFSE FI while the last term on the right represents rate of production of new cells due to cell division. A derivation of this model following the mass conservation principles of Sinko-Streifer [97] or Bell-Anderson [26] can be found in the Appendix of [21]. As noted in Chapter 1, the dependence of the functions α and β on the structure variable z is to be understood as a surrogate for the division dependence of the proliferation and death rates (that is, it is not believed that CFSE FI is a causative factor in changing the rates).

Because flow cytometry data is typically represented in a logarithmic scale, it is convenient to make the change of variables $z = \log_{10} x$. The resulting model when $v(x) = -cx$ (which was assumed in

[21, 77]) and $\tilde{n}(t,z) \equiv n(t,10^z)$ is

$$\frac{\partial \tilde{n}}{\partial t}(t,z) + \frac{\partial[\tilde{v}(z)\tilde{n}(t,z)]}{\partial z} = -(\tilde{\alpha}(z) + \tilde{\beta}(z))\tilde{n}(t,z) + \chi_{[z_{\min}, z_{\max} - \log_{10}\gamma]} 2\gamma\tilde{\alpha}(z + \log_{10}\gamma)\tilde{n}(t, z + \log_{10}\gamma), \tag{2.2}$$

where $\tilde{v}(z) = -\tilde{c} = -c/ln10$, and $\tilde{\alpha} and \tilde{\beta}$ are the appropriately defined *effective* cell proliferation and death rates, respectively. The initial and boundary conditions for the model are

$$\tilde{n}(0,z) = \tilde{\Phi}(z)$$
$$\tilde{n}(t, z_{\max}) = 0.$$

In subsequent discussions, the tildes on the parameters α, β, c, v and the states n in (2.2) will be dropped. Some key tacit assumptions of this model are:

(i) Division numbers are strongly correlated with FI;

(ii) FI is proportional to total CFSE mass;

(iii) Total CFSE is divided equally among daughter cells with each division;

(iv) The rate of label loss $v(z)$, the proliferation rate $\alpha(z)$, and the death rate $\beta(z)$ do not depend on time.

Given assumptions (i) - (iii), assumption (iv) is equivalent to stating that birth, death, and label loss rates depend only on division number and thus that these rates can be determined as functions of label intensity z. In an effort to improve the model (2.2), the validity of these assumptions must be investigated and some assumptions modified in their interpretation and implementation to produce significantly improved model agreement with the data. To this end, the CFSE data described in Chapter 1 was used in order to estimate the functions $\alpha(z)$, $\beta(z)$ (which were parameterized as piecewise linear functions) as well as the parameters c and γ in an Ordinary Least Squares (OLS) framework. A detailed discussion of the inverse problem procedure can be found in [21]. Such a discussion is omitted here (although Section 2.2.5 is similar). Instead, a brief summary of the results obtained in [21] is provided in order to motivate the improvements to the model (2.2) discussed in the remainder of this chapter.

The OLS best-fit solution using the original model (2.2) is shown in comparison to the data in Figure 2.1. It is clear that this model is insufficient to describe the dynamics observed in the data set used to calibrate the model. However, by allowing the proliferation rate to be explicitly time dependent (that is, $\alpha = \alpha(t,z)$ rather than $\alpha = \alpha(z)$), the model could be much more accurately fit to the data. Using statistical tests for model refinements based upon the residual-sum-of-squares (RSS) [10, 11, 16, 22, and references therein] it was shown that the improved accuracy was more than justified by the increase in parameters required to parameterize α as a function of time as well as measured FI.

Because one of the primary goals of the mathematical modeling of flow cytometry data is the determination of the manner in which proliferation and death rates change with division number, additional model improvements were hypothesized and tested. In particular, it was found that the dependence of

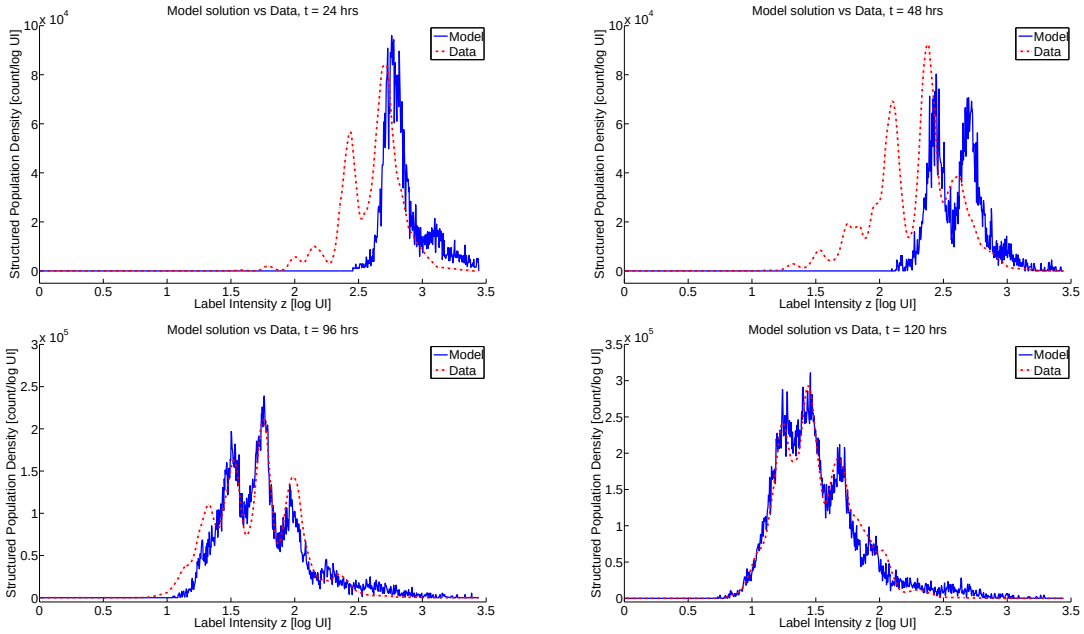

Figure 2.1: OLS best-fit solution of the original PDE model (2.2) in comparison to the data at $t = 24$, 48, 96, and 120 hours; from [21].

the death rate function β on measured FI z (and hence, on division number) must be maintained in order to accurately fit the data. (That is, a single, constant death rate does not provide a satisfactory fit of the data.) It was also shown that a 'translated coordinate' $s = z + ct$, by correcting for the slow natural decrease of CFSE FI over time, was more strongly correlated with division number than the original structure variable z. As such, the functions $\alpha(t, s)$ and $\beta(s)$ provided an estimate of the proliferation and death rates, respectively, which more accurately indicated division-linked changes in those rates.

With these improvements, the original model (2.2) could be rewritten as

$$\begin{aligned}\frac{\partial n(t,z)}{\partial t} + \frac{\partial[-cn(t,z)]}{\partial z} &= -(\alpha(t, z+ct) + \beta(z+ct))n(t,z) \\ &+ \chi_{[z_{\min}, z_{\max} - \log\gamma]} 2\gamma\alpha(t, z + ct + \log\gamma)n(t, z + \log\gamma).\end{aligned} \tag{2.3}$$

The OLS best-fit estimate for this model is shown in comparison to the data in Figure 2.2. It is clear that the improved model very closely fits the data. Yet there are multiple issues which must be explained. For instance, the 'translated coordinate', while heuristically motivated (see [21]) is not intuitive and needs mathematical explanation. It turns out that this improvement can be explained in the language of mechanics as a consequence of a change to a moving coordinate system (Section 2.2.5). This change in coordinate system will depend upon the manner in which CFSE FI is naturally lost from the measured population of cells. While the current model assumes an exponential rate of decay ($dx/dt = -cx$), a more accurate model of label loss can be determined by analyzing additional data sets (Section 2.2.3) to show that the rate of exponential decrease itself decreases in time.

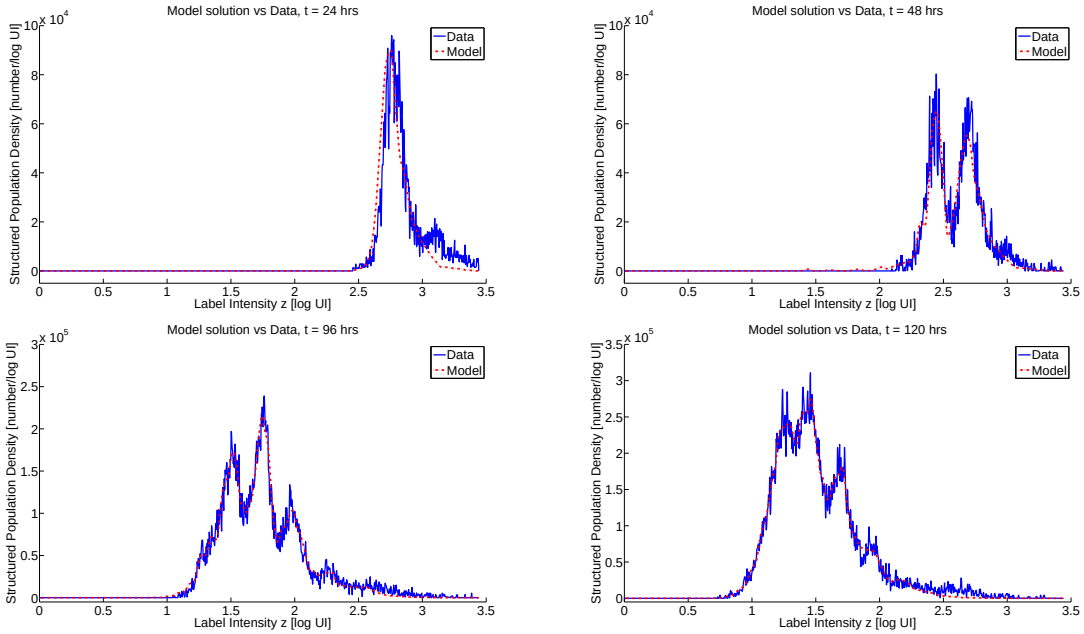

Figure 2.2: OLS best-fit solution of the improved PDE model (2.3) in comparison to the data at $t = 24$, 48, 96, and 120 hours; from [21].

Most importantly, there is a physiological impossibility in the model (2.3) which must be addressed. The parameter γ is used in [21] to represent an unknown process responsible for determining the CFSE FI of two daughter cells given the CFSE FI of a mother cell, without regard for what that process might be (see also [77] where the parameter γ was first introduced). Mathematically, it follows from the form of Equation (2.3) that the parameter γ determines the peak-to-peak separation between subsequent generations of cells (i.e., each generation has a CFSE FI approximately $\log_{10} \gamma$ less than the previous generation in the log FI coordinate z). Given the definition of γ as the ratio of CFSE FI of a mother cell to that of a daughter cell as well as the assumption that CFSE FI is a mass-like quantity, it is expected that $\gamma \geq 2$, with $\gamma > 2$ if label is lost during the process of cell division. However, the best fit parameter from [21] was $\gamma^* = 1.5169$, implying the creation of label at division. Similar results were also obtained in [77]. Indeed, forward simulations of the above model demonstrate that $\gamma = 2$ is significantly too large to fit the given data set, regardless of the values assigned to other parameters. Some work is required to make the model (2.3) more physically relevant by explaining the mechanism underlying the parameter γ and, in doing so, resolving an apparent physical impossibility.

2.2 Model Revisions and Improvements

Presented now are several significant revisions and clarifications of the model (2.3). First is the replacement of the parameter γ with an actual physiological mechanism responsible for the dilution of CFSE FI when a cell divides. A derivation of the PDE model is then presented which includes this mechanism, as

well as unspecific functions $v(t,x)$, $\alpha(t,x)$, and $\beta(t,x)$ for the label loss rate, proliferation rate, and death rate, respectively. Using two additional data sets, it is shown that the label loss rate is best described by a Gompertz decay function. Finally, in keeping with the goal of examining the division dependence of the proliferation and death rates, possible parameterizations of these two functions are discussed at length and a mathematical motivation of the 'translated coordinate' is presented.

2.2.1 Physiological Interpretation of γ

As stated above, it follows from the assumptions of the model that $\gamma < 2$ is a physical impossibility. Meanwhile, $\gamma \geq 2$ results in significantly too much peak-to-peak separation in the model solution when compared to the data. One possible solution conjectured in [21] to explain this discrepancy was that the measurement of CFSE FI may be indicative of the concentration of CFSE, rather than its mass. The observations, then, would represent an effective integration over the various cell volumes present in the data. While mathematically appealing, this explanation does not appear to be the case. Physically, one expects that measured CFSE FI would depend on the number of CFSE molecules within the cell, and hence on the mass of CFSE. Indeed, when cells stained with CFSE are introduced to a stimulating agent, the cells quickly increase in size (thus decreasing the CFSE concentration), but the measured CFSE FI is essentially unchanged [82, Fig. 6].

We propose here an alternative solution to this apparent γ-related dilemma. While it is often stated that the subsequent peaks in the CFSE histogram data are evenly spaced [80], close observation reveals that this is not actually the case. Although the peaks corresponding to low division numbers (Generations 0, 1, 2, 3) are approximately evenly spaced, peaks corresponding to larger division numbers are closer and closer together [82, Fig. 1]. In other words, the parameter γ appears to change with division number.

These observations can be collectively explained by the consideration of cellular autofluorescence and its effects on the measurement process. As discussed in Chapter 1, the flow cytometer measurement process uses light as a surrogate for CFSE content. However, not all light incident upon the photodetector is the result of emission from CFSE molecules. All cells, even those unstained with CFSE, have a natural brightness and will give off small but detectable amounts of light. We assume here that this feature, the *cellular autofluorescence* does not change as cells divide and does not decay slowly like CFSE fluorescence. It may vary with time for other reasons [3], but we ignore this in our initial treatment.

Let X_i be the total measured FI of a single cell, measured when that cell has undergone i divisions. (The use of the capital letter is meant to distinguish this discrete quantity from the continuous state variable to be used in the revised model derived below.) Under the assumption that cellular autofluorescence intensity (AutoFI) and CFSE FI are additive [64], the total fluorescence intensity of a cell is

$$X_i = X_i^{\mathrm{CFSE}} + X^{\mathrm{Auto}}. \tag{2.4}$$

Because AutoFI does not change as a cell divides, it follows that this cell with intensity X_i will generate two cells in the next generation, each of which has total FI

$$X_{i+1} = X_i^{\mathrm{CFSE}}/2 + X^{\mathrm{Auto}}. \tag{2.5}$$

Contrary to previous interpretations of the parameter γ, one can see from Equation (2.5) that it is actually expected that the ratio of total FI of a mother cell to that of a daughter cell is expected to be less than 2. Moreover, provided $X_i^{\text{CFSE}} >> X^{\text{Auto}}$, this ratio is approximately equal to two. With each division, X_i^{CFSE} decreases and the ratio decreases; as X^{Auto} accounts for a larger and larger percentage of the total measured FI, the ratio decreases quicker and quicker until $X_i^{\text{CFSE}} \approx 0$ and the ratio of mother-to-daughter intensities is approximately 1. Thus it appears that cellular autofluorescence is sufficient to account for the observed relationships between subsequent division peaks in the data. Indeed, this is shown to be the case in Section 2.4.

We remark that the phenomenon of autofluorescence when using fluorescent dyes to study biological materials is not particularly new. In fact, the role of AutoFI described above was acknowledged specifically for CFSE data sets as early as 1996 [59], and a formula corresponding to (2.5) above appears in [80]. However, autofluorescence has not been used in previous PDE formulations [21, 75, 77] to describe the dilution of CFSE by division. Thus the incorporation of AutoFI into the mathematical model presented below is an important revision to the physiological basis of the PDE model.

2.2.2 Revised Model Derivation

At this point, we pause momentarily in order to revise the PDE model derivation from the Appendix of [21]. We do so to provide a general framework with which to build additional improvements to the model and inverse problem procedure. As before, the derivation follows the mass-balance principles of the Bell-Anderson [26] and Sinko-Streifer [97] models.

Let $n(t, x)$ be the structured population density (cells/UI) of a population of cells labeled with CFSE, where t is time (in hours) and the structure variable x is the fluorescence intensity of a cell. Then

$$N(t) = \int_{x_0}^{x_1} n(t, x)dx. \tag{2.6}$$

represents the total number of cells with fluorescence intensity in (x_0, x_1) at time t. Here x_0 and x_1 are arbitrary. Let $\Delta x(t, x, \Delta t)$ be the average increase of FI of cells with initial intensity x during the interval $(t, t + \Delta t)$ and assume that Δt is chosen such that $|\Delta x| << x_1 - x_0$ (so that the number of cells which move into the region via division and subsequently divide, die, or drift out of the region is negligible). It should be noted that Δx will be non-positive (as cells cannot increase in FI). Thus subtraction by Δx actually results in a larger value. While counterintuitive, this definition is maintained in order to harmonize with other structured population models.

Consider the change in $N(t)$ during the time interval $(t, t + \Delta t)$, i.e., the quantity $N(t + \Delta t) - N(t)$. Five possible contributions will be considered:

(i.) Cells with intensity in the interval $[x_1, x_1 - \Delta x(t, x_1, \Delta t)]$, losing FI according to Δx:

$$\int_{x_1}^{x_1 - \Delta x(t, x_1, \Delta t)} n(t, x)dx.$$

(ii.) Cells with intensity in the interval $[x_0, x_0 - \Delta x(t, x_0, \Delta t)]$, losing FI according to Δx:

$$\int_{x_0}^{x_0 - \Delta x(t,x_0,\Delta t)} n(t,x)dx.$$

(iii.) Cells which would have contributed to $N(t + \Delta t)$ had they not died:

$$\int_t^{t+\Delta t} \int_{x_0 - \Delta x(t,x_0,t+\Delta t-\tau)}^{x_1 - \Delta x(t,x_1,t+\Delta t-\tau)} \beta(\tau,x)n(\tau,x)dxd\tau.$$

(iv.) The disappearance of cells from the region due to proliferation:

$$\int_t^{t+\Delta t} \int_{x_0 - \Delta x(t,x_0,t+\Delta t-\tau)}^{x_1 - \Delta x(t,x_1,t+\Delta t-\tau)} \alpha(\tau,x)n(\tau,x)dxd\tau.$$

(v.) The gain of daughter cells (two of them) in the region as a result of proliferation in the parent region:

$$\chi_{[x_a,x^*]} 2 \int_t^{t+\Delta t} \int_{2(x_0 - \Delta x(t,x_0,t+\Delta t-\tau))-x_a}^{2(x_1 - \Delta x(t,x_1,t+\Delta t-\tau))-x_a} \alpha(\tau,x)n(\tau,x)dxd\tau,$$

where $x^* = x_{\max}/2 + x_a$ and x_a is the natural autofluorescence of unstained cells.

We remark that α and β are the rates of cell proliferation and death, respectively, with units hr^{-1}. It follows that the difference $N(t + \Delta t) - N(t)$ is the sum of the components (i.) and (v.) less the contributions of components (ii.) - (iv.). Following the standard procedure of dividing by Δt and letting $\Delta t \to 0$, we obtain $\frac{dN}{dt}$ on the left side of the equation. We now treat the right side of the equation term by term.

For the first term on the right side, if $n(t, x)$ is continuous in t and x (a reasonable assumption), the mean value theorem (MVT) implies that there exists a $\theta \in [x_1, x_1 - \Delta x(t, x_1, \Delta t)]$ such that

$$\int_{x_1}^{x_1 - \Delta x(t,x_1,\Delta t)} n(t,x)dx = -\Delta x(t,x_1,\Delta t)n(t,\theta).$$

Assuming Δx is continuous in Δt (that is, there is no instantaneous label loss) and varies smoothly,

$$\lim_{\Delta t \to 0} \frac{-\Delta x(t,x_1,\Delta t)}{\Delta t} n(t,\theta) = -v(t,x_1)n(t,x_1). \tag{2.7}$$

where we have defined $\frac{dx}{dt} = v(t, x)$, the instantaneous rate of FI change of cells with intensity x and time t. Applying the same argument for the second term,

$$-\int_{x_0}^{x_0 - \Delta x(t,x_0,\Delta t)} n(t,x)dx = v(t,x_0)n(t,x_0). \tag{2.8}$$

In the consideration of the third term, define

$$u_\beta(\tau) = \int_{x_0-\Delta x(t,x_0,t+\Delta t-\tau)}^{x_1-\Delta x(t,x_1,t+\Delta t-\tau)} \beta(\tau,x)n(\tau,x)dx.$$

Then if $\Delta x(t,x,\Delta t)$ and $\beta(\tau,x)n(t,x)$ are continuous functions of their variables, so is $u_\beta(\tau)$ and by the MVT, there exists a $\theta' \in [t, t+\Delta t]$ such that

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} u_\beta(\tau)d\tau = u_\beta(\theta').$$

Thus it follows that

$$\lim_{\Delta t\to 0} u_\beta(\theta') = u_\beta(t) = \int_{x_0}^{x_1} \beta(t,x)n(t,x)dx, \tag{2.9}$$

assuming $\Delta x(t,x,0) = 0$ for all t, x (which follows from the previous assertion regarding the smoothness of Δx in Δt). Using a similar argument for the fourth term,

$$\lim_{\Delta t\to 0} u_\alpha(\theta') = u_\alpha(t) = \int_{x_0}^{x_1} \alpha(t,x)n(t,x)dx, \tag{2.10}$$

where $u_\alpha(\tau)$ has the obvious definition. For the final term, the same argument along with the change of variables $\xi = (x + x_a)/2$ results in

$$\chi_{[x_a,x^*]}2\lim_{\Delta t\to 0} u_{\tilde{\alpha}}(\theta') = \chi_{[x_a,x^*]}4\int_{x_0}^{x_1} \alpha(t,2x-x_a)n(t,2x-x_a)dx. \tag{2.11}$$

Altogether, we can assemble (2.7) - (2.11) to obtain

$$\begin{aligned}\frac{dN}{dt} &= -v(t,x_1)n(t,x_1) + v(t,x_0)n(t,x_0) - \int_{x_0}^{x_1}\beta(t,x)n(t,x)dx \\ &\quad - \int_{x_0}^{x_1}\alpha(t,x)n(t,x)dx + \chi_{[x_a,x^*]}4\int_{x_0}^{x_1}\alpha(t,2x-x_a)n(t,2x-x_a)dx.\end{aligned}$$

On the left side, differentiating $N(t) = \int_{x_0}^{x_1} n(t,x)dx$ with respect to t results in

$$\frac{dN}{dt} = \int_{x_0}^{x_1}\frac{\partial n(t,x)}{\partial t}dx.$$

Finally, by applying the Fundamental Theorem of Calculus to the first two terms on the right side, simplifying and rearranging,

$$\begin{aligned}&\int_{x_0}^{x_1}\frac{\partial n(t,x)}{\partial t} + \int_{x_0}^{x_1}\frac{\partial(v(t,x)n(t,x))}{\partial x} = \\ &- \int_{x_0}^{x_1}(\alpha(t,x)+\beta(t,x))n(t,x) + \chi_{[x_a,x^*]}4\int_{x_0}^{x_1}\alpha(t,2x-x_a)n(t,2x-x_a).\end{aligned}$$

Equivalently (because x_0 and x_1 were arbitrary),

$$\begin{aligned} \frac{\partial n(t,x)}{\partial t} &+ \frac{\partial [v(t,x)n(t,x)]}{\partial x} = \\ &- (\alpha(t,x)+\beta(t,x))n(t,x) + \chi_{[x_a,x^*]} 4\alpha(t,2x-x_a)n(t,2x-x_a). \end{aligned} \tag{2.12}$$

At the right boundary ($x = x_{\max}$), we expect that there are no cells which can drift (via label loss) into the computational domain. At the left boundary, a zero flux condition is imposed to prevent cells from drifting to CFSE FI values less than the AutoFI of unlabeled cells. Thus the boundary conditions are

$$n(t, x_{\max}) = 0 \tag{2.13}$$

$$v(t, x_a)n(t, x_a) = 0. \tag{2.14}$$

In general (see Equation 2.19) the label loss function will satisfy $v(t, x_a) = 0$ for all t, and hence the left boundary condition will be trivially satisfied. Finally, we assume we are given some initial condition

$$n(0, x) = \Phi(x), \tag{2.15}$$

which is the initial distribution of cells as a function of FI.

We remark that the above derivation is not very different from that already presented in [21]. The key differences are the notational change in permitting the dependence of the proliferation and death rates (α and β) and the label loss rate (v) on both time t and measured FI x. In addition to accounting for the presence of cellular AutoFI, this model also explicitly incorporates the even division of CFSE between daughter cells.

2.2.3 Gompertz Decay of Label

Given the model (2.12), we now turn our attention to the label loss rate $v(t, x)$. Because the mathematical model estimates cell proliferation and death rates in terms of the CFSE FI structure variable (as a surrogate for division number), the manner in which CFSE naturally decays directly affects the cell turnover parameter estimates. (This is particularly true when using the 'translated variable'; see Section 2.2.5). Thus, our understanding of the underlying biology (in the form of cell proliferation and death rate estimates) is closely tied to the accurate modeling of label decay. In order to provide parameter estimates which are unbiased, it is of vital importance that the label loss rate $v(t, x)$ accurately reproduces the natural decrease in CFSE FI observed in the data. Previous authors have also highlighted an accurate model of label decay as a necessary factor in describing CFSE data sets [61, 64].

It was hypothesized in [77] that an exponential rate of loss is sufficient to model the label loss observed in the data, and this assumption was incorporated into Equation (2.1). In order to test this assumption, a PBMC culture was taken from two donors and stained with CFSE following the standard procedure. However, these cells were not stimulated to divide. Because only viable cells are included when the cytometry data is gated, any decrease in mean FI in the population must be the result of natural CFSE FI decay. Over the course of 160 hours, cells from each donor were measured at 24 distinct time points

in triplicate and the mean total FI of each sample was recorded. The data can be found in Table 2.1 and is shown graphically in Figure 2.3.

We would like to determine what functional forms might be used in order to quantify the label loss observed in the data. Following the assumptions of [21, 77], we begin with a model of label loss that decays exponentially to the autofluorescence of unlabeled cells,

$$x_1(t) = (x(0) - x_a)e^{-ct} + x_a. \tag{2.16}$$

However, it appears from the data (particularly for Donor 1) that the rate of exponential decay of label may itself decrease as a function of time. This can be modeled by the Gompertz decay process [69, pg. 12]

$$x_2(t) = (x(0) - x_a)\exp\left(-\frac{c}{k}\left(1 - e^{-kt}\right)\right) + x_a. \tag{2.17}$$

The loss rate function, of vital importance to the PDE formulation (2.12), is $v(t,x) = \frac{dx}{dt}$. Thus the equations (2.16) - (2.17) correspond to the loss rate functions

$$v_1(x) = -c(x - x_a), \tag{2.18}$$

and

$$v_2(t,x) = -c(x - x_a)e^{-kt}, \tag{2.19}$$

respectively. We remark that (2.17) is a generalization of (2.16), the latter being the limiting value (as $k \to 0$) of the former. Thus, it would be ideal to fit both models to the data and use statistical tests to determine if the model refinement is warranted. However, it is not possible to uniquely identify the parameter x_a in either of the two models using an ordinary least squares procedure with only the data from Table 2.1. On the other hand, the parameter x_a appears in other parts of the model (2.12), and it seems possible to conclude that this parameter would be readily identifiable once incorporated into (2.12) with the full proliferation assay data.

In order to at least begin to compare these two models, we set the parameter x_a to the physiologically reasonable value of 50 in both models. (Values reported in the literature ranged from 10 to 100 [79, Figure 1], [82, Figure 2] and [92, Figure 3].) An ordinary least squares (OLS) procedure is then used to fit the remaining parameters (c and $x(0)$ for the exponential model, c, k and $x(0)$ for the Gompertz model) for data from both donors. The results for the exponential model are shown in Figure 2.4 and the results for the Gompertz model are shown in Figure 2.5. Based upon the lines of best fit in comparison with the data, it seems clear that the Gompertz decay model more accurately describes the available data. This is confirmed by the comparison of cost given in Table 2.2.

A few additional comments are in order. First, we remark that, while the parameter $x(0)$ is included in the models (2.16) and (2.17) used to fit the CFSE label loss data sets from Table 2.1, it is actually the loss rate function (either (2.18) or (2.19)) that will be included in the PDE model (2.12) and thus this parameter will not actually require estimation in the full PDE model. Also, when using CFSE-labeled cells for a proliferation assay, there is at minimum an additional hour of preparation time (so that the cells can be stimulated to divide) between CFSE labeling and the first measurement time. Thus we

Table 2.1: Mean CFSE FI for unstimulated cells rounded to the nearest integer. Data collected from Donor 1 and Donor 2. Several outliers are noticeable in data from Donor 2 and have been marked with an asterisk. All data is given in arbitrary units of intensity (UI).

Time (hours)	Donor 1			Donor 2		
0.00	44311	43272	45369	40878	41593	41993
2.00	33782	37720	36961	36755	32585	25705*
4.00	29331	30043	29634	30818	28565	27144
6.00	29235	28526	31283	26498	27666	26354
8.00	25899	27229	29839	25856	18404*	25060
10.00	26651	27691	27406	24846	25336	–
12.50	29471	27610	27852	24172	25506	26290
19.25	27201	27254	24718	30272*	26922	26937
21.25	27062	23601	25342	27436	27646	29527
23.25	20758	24967	24640	25680	26474	26482
25.25	26585	23512	23400	25947	26711	26379
27.25	23356	24882	22898	26040	25551	28393
29.25	23660	21729	24288	23975	22490	23471
31.50	22768	20914	21268	21153	21244	21759
33.25	23897	24198	24758	25337	24053	24749
50.25	22623	23504	24696	26138	25672	27361
54.75	21910	21120	21986	25777	24564	26069
59.25	21877	26290	23829	21932	21302	27558*
77.25	22099	24160	21420	25769	24108	26511
85.75	20731	21827	22108	26604	26293	26306
98.75	21468	20993	21220	22869	21760	22579
123.75	18836	18887	18313	20369	20533	21003
131.25	20132	20119	20799	20908	22234	26666*
150.75	22199	19822	–	26542	23417	–

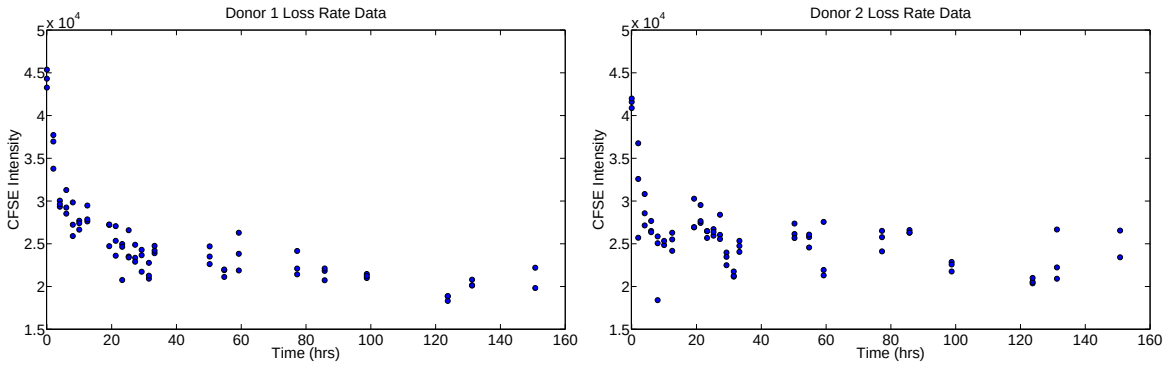

Figure 2.3: Mean CFSE FI data for unstimulated cells from donor 1 (left) and donor 2 (right).

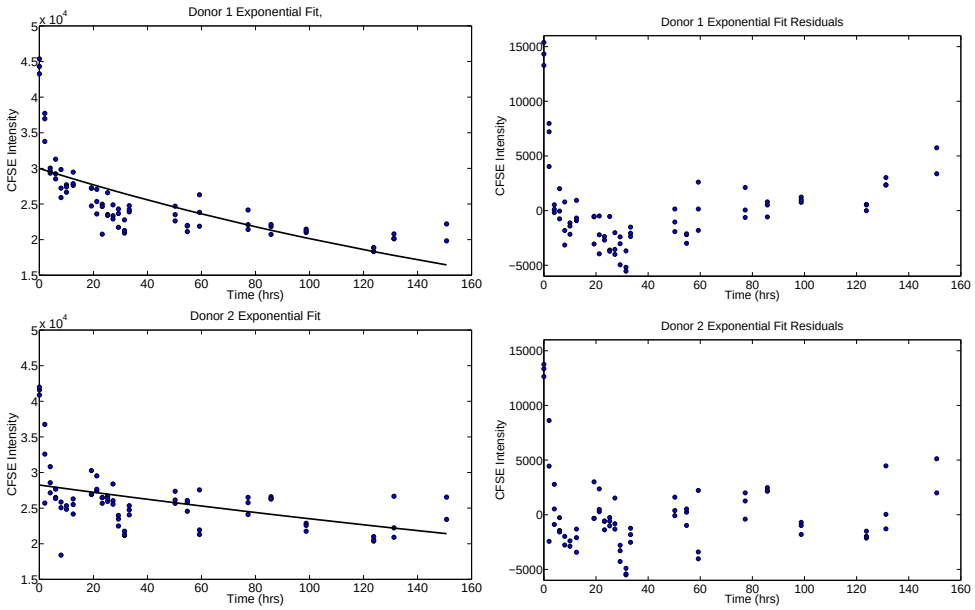

Figure 2.4: Results of fitting the exponential model (2.16) to the mean CFSE data in Table 2.1. For both Donor 1 (top) and Donor 2 (bottom), we see from both the fit to the data (left) and the residual plot (right) that the model is not capable of accurately replicating the observed data (when x_a is set to 50 UI).

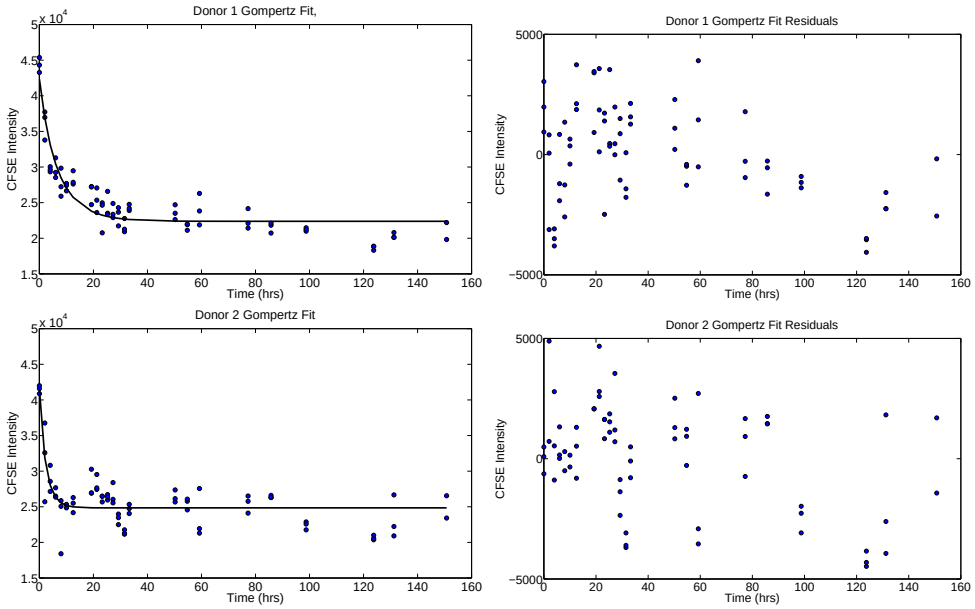

Figure 2.5: Results of fitting the Gompertz model (2.17) to the mean CFSE data in Table 2.1. The fit of the model to the data is much improved (when compared to the exponential fit in Figure 2.4) for both donors (when x_a is set to 50 UI).

Table 2.2: OLS cost values for fitting the exponential (2.16) and Gompertz (2.17) models to the mean CFSE FI data for unstimulated cells from two donors (Table 2.1). While model comparison tests are not applicable because of a lack of parameter identifiability, there is a clear reduction in cost when using the Gompertz model, as well as a clear improvement in the model fit (compare Figures 2.4 and 2.5).

Model	Donor 1	Donor 2
Exponential	1.176515×10^9	1.073325×10^9
Gompertz	2.853192×10^8	4.285324×10^8

would expect slightly different parameter estimates when either of these models is used with proliferation assay data. Moreover, because this additional hour of preparation time occurs during the first hour after staining, the rapid decay observed in both data sets for $0 \leq t \leq 5$ may be partially removed from the data. In fact, the estimation of the label loss function (either exponential or Gompertz) is highly sensitive to the length of time separating initial labeling from stimulation. In some experiments, this sensitivity is minimized by separating these two events by a full 24 hours [84]. Because of these considerations, it is unclear exactly how much improvement to expect (in additional data sets) from using (2.19) as opposed to (2.18) in (2.12), and the magnitude of the improvement will probably depend upon the (unknown) length of time between staining and stimulation.

We also remark that the primary reason for the failure of the exponential model is the location of the equilibrium point in the data (as compared to that predicted by the model). We see from Equation (2.16) that the exponential model predicts $x \to x_a$ as $t \to \infty$. However, it is known that CFSE stained cells retain detectable fluorescence for up to several weeks in vivo [82]. The exponential model cannot accurately account for both the rapid decline in CFSE FI during the first few hours of staining and the slow decline once the label has been stably incorporated. (We can, for the current data sets, allow x_a to attain values of approximately 2.2×10^4; in this case the exponential model fits the data at least as well as the Gompertz model. However, this is not a physiologically reasonable value of x_a.)

Physiologically , it is known that after the conversion of CFDA-SE to CFSE by intracellular esterases, CFSE can still exit the cell at a slow rate (compared to free diffusion). However, the succinimidyl group of CFSE reacts covalently with amines attached to intracellular proteins. While some of the resulting conjugates are short-lived (either because they exit the cell or are rapidly degraded) other conjugates are stably incorporated inside the cell and remain so for an extended period of time. These stable conjugates are decreased further only by the natural turnover of the intracellular proteins to which they are bound [89]. These processes combine to produce the commonly observed "biphasic decay" of CFSE FI over time [84, 104]. In other words, it seems necessary for the rate of CFSE FI (exponential) decay to decrease in time. This is precisely the feature of the Gompertz decay model [69].

2.2.4 Model Change of Variables

Before proceeding to a consideration of possible parameterizations of the proliferation and death rate functions α and β, we pause briefly to consider a change of variables which will aid in the calculation of numerical solutions. While the model (2.12) (with $v(t, x)$ given by (2.19)) and its associated initial and boundary conditions suitably describe the dynamics for a CFSE-labeled population of dividing

lymphocytes, the model is not conducive to finite difference methods for numerical solutions. The CFL condition for stability requires

$$\Delta t < \frac{\Delta x}{\max |v(t,x)|} = \frac{\Delta x}{\max c(x - x_a)}. \tag{2.20}$$

Because $x_{\max} >> x_a$, the computational domain is quite large, and $\max(x - x_a) \sim 10^4$. Moreover, this large domain must have a relatively fine mesh, as features of the solution become less distinguishable with increasing division number. We see that, given $c \sim 0.1$ and $\Delta x \sim 0.1$, it would take upward of 10^5 time steps to compute the solution out to $t = 120$ hours, and this must be done for 10^5 points on the structure variable grid. The resulting computations can take in excess of several hours running in MATLAB on a desktop machine.

Rather than attempt these expensive computations, we seek a change of variables that will lead to a faster numerical solution. The most immediate choice is to use the change of variables $z = \log_{10} x$, as the data is given in this coordinate. While this change of variables was effective in [21, 77], it is less effective here because of the different form of the label loss rate function $v(t, x)$. Instead, we use the change of variables $y = \log_{10}(x - x_a)$. Then $x = 10^y + x_a$ and

$$\frac{dy}{dx} = \frac{1}{(x - x_a)\ln(10)}.$$

Let $\tilde{n}(t, y) = 10^y \ln(10) n(t, x(y)) = 10^y \ln(10) n(t, 10^y + x_a)$. We remark that the factor $10^y \ln(10)$ arises from the chain rule in the integral form of (2.12) and is needed to conserve the total label in the population. With this change of variables the new PDE model is

$$\begin{aligned} \frac{\partial \tilde{n}}{\partial t} - ce^{-kt}\frac{\partial}{\partial y}\left[\frac{\tilde{n}(t,y)}{\ln 10}\right] &= -(\tilde{\alpha}(t,y) + \tilde{\beta}(t,y) - ce^{-kt})\tilde{n}(t,y) \\ &+ \chi_{(-\infty,y^*]} 2\tilde{\alpha}(t, y + \log_{10} 2)\tilde{n}(t, y + \log_{10} 2), \end{aligned} \tag{2.21}$$

where $\tilde{\alpha}(t, y) = \alpha(t, x(y))$, $\tilde{\beta}(t, y) = \beta(t, x(y))$ and $y^* = y_{\max} - \log_{10} 2$. The new initial condition is $\tilde{\Phi}(y) = 10^y \ln(10)\Phi(t, 10^y + x_a)$. The right boundary condition follows immediately from (2.13) while the left boundary has been removed to $y = -\infty$. We remark that the CFL condition for the PDE (2.21) is

$$\Delta t < \frac{\Delta y \ln 10}{ce^{-kt}}$$

which is significantly easier to satisfy. The change of variables $y = \log_{10}(x - x_a)$ is a parameter-dependent change of variables, technically requiring a 're-gridding' of the solution each time the parameters (specifically x_a) are changed in an optimization routine for the inverse problem. Because we use a time-stepping finite-difference method to compute the forward solution for a given set of parameters, this requirement does not constitute a great computational setback. In fact, observation of (2.21) reveals that the parameter x_a does not appear directly in the equation to be solved, only in the change of variables that gives rise to the equation.

In the remainder of this chapter, the tildes over the functions α, β, n and Φ will be ignored. As has already been noted, the dependence of the functions α and β on the structure variable (either x or

y) is intended only as a surrogate for the dependence of the proliferation and death rates, respectively, on division number (see Section 2.4). As such, the distinction between $\alpha(t,x)$ and $\tilde{\alpha}(t,y)$ (or $\beta(x)$ and $\tilde{\beta}(y)$) is of minimal importance–the derivation of the 'translated variable' below applies equally as well to one as the other.

2.2.5 Parameterizations of α and β

We next turn our attention to the parameterizations of the functions $\alpha(t,x)$ and $\beta(t,x)$. Because our goal is the estimation of lymphocyte division and death rates from data, we use finite-dimensional approximations of the function spaces containing α and β so that the problem is computationally tractable and theoretically sound [16, 17]. Previous work has established that division-linked changes in proliferation and death rates are an important aspect of an accurate mathematical model [21, 72, 76, 77]. One of the primary motivating assumptions behind the use of a PDE model for fitting CFSE data is that the dilution of CFSE dye by division allows for the structure variable (in this case, y) to be used as a surrogate for division number [21, 75, 77]. Thus division-dependent changes in proliferation and death rates are encapsulated in the structure dependence of the functions α and β.

While a straightforward implementation of structure dependence for α and β has proven effective, the fact that the measured FI of a cell slowly decreases in time as a result of label loss indicates that one should take care in accounting for the correlation between division number and structure variable. As discussed at greater length in [7, 21], this label loss causes such correlation to lessen significantly. Alternatively, one might consider the total FI that *would* have been measured for a cell, if that cell did not experience any label loss. The key argument is to identify a cell not by its current state y, but rather by the state it would have in the event it did not undergo label loss. We begin by demonstrating that such a 'translated coordinate' is mathematically equivalent to deriving a model in terms of an ideal label which does not decay and then changing one's frame of reference to a moving coordinate system in which the label does appear to decay (i.e., the one relative to which the data is actually taken). After this derivation, several possible parameterizations are considered for the proliferation and death rate functions α and β.

The 'translated variable' as a change of coordinate systems

Consider a population of cells labeled with a (hypothetical) fluorescent dye which *does not decay* as the cells grow, divide, and die. Let s be the measured FI (in units of intensity, UI, as before) of such cells and let $n(t,s)$ be the structured population density (cells/UI) of cells with FI s at time t. Then for a given FI s, the population density is governed by a rate equation which can be written as

$$\frac{dn}{dt}(t,s) = -\left(\alpha(t,s)+\beta(t,s)\right)n(t,s) + \chi_{[s_a,s^*]}4\alpha(t,2s-s_a)n(t,2s-s_a), \tag{2.22}$$

where s^* and s_a are defined as in Section 2.2.2. A derivation of this equation follows immediately from the mass-conservation principles of [26, 97] and Section 2.2.2 above. One can now make the change of

variables $\tilde{s} = \log_{10}(s - s_a)$ as in Section 2.2.4 to obtain

$$\frac{d}{dt}\left[10^{\tilde{s}}\log(10)n(t,\tilde{s})\right] = -\left(\tilde{\alpha}(t,\tilde{s}) + \tilde{\beta}(t,\tilde{s})\right)\left(10^{\tilde{s}}\log(10)n(t,\tilde{s})\right) + \chi_{(-\infty,\tilde{s}^*]}2\tilde{\alpha}(t,\tilde{s}+\log_{10}2)\left(10^{\tilde{s}}\log(10)n(t,\tilde{s}+\log_{10}2)\right). \quad (2.23)$$

Typically, when making a change of variables in differential form, one would simply divide through by the quantity $10^{\tilde{s}}\log(10)$, which is always nonzero. However, because we will be making a time-dependent change of coordinates below, this induces a time dependence in the variable $\tilde{s}$ (when viewed from the new frame of reference). As such, misleading results can be obtained if the factors $10^{\tilde{s}}\log(10)$ are eliminated from the above equation.

While Equation (2.23) would be adequate in the situation for which total FI remains constant (unless diluted by division or removed by cell death) it has already been acknowledged that CFSE (along with essentially all other intracellular dyes) naturally degrades and/or is turned over until FI is no longer discernible from background autofluorescence. Thus it is of interest to introduce a natural label loss velocity v to account for the changing label intensity. That is, the reported intensity coordinate system is actually changing in time (i.e., a moving coordinate system with velocity v). We therefore introduce a change of coordinates involving this velocity to decouple the degradation process from the cell division/proliferation/death processes. Moreover, since the data is recorded relative to this changing structure variable, it is useful to view the proliferation and death rates *relative to this new coordinate system.*

Although not common in the biological sciences, it is altogether common in the physical sciences and engineering to consider velocities (i.e., rates of change) relative to different coordinate or reference frames. For example, in the mechanics and motion of continua (elasticity and fluids) and deformable bodies [14, 47, 48, 83, 88], it is frequent to encounter velocities relative to a *fixed* coordinate system (in a *Lagrangian* formulation) or relative to a *moving* coordinate system (in an *Eulerian* formulation). One description for motion is made in terms of the material or fixed referential coordinates, and is called a material description or the *Lagrangian description.* In this formulation, an observer standing in the fixed referential frame observes the changes in the position and physical properties as the material body moves in space as time progresses. This formulation focuses on individual particles as they move through space and time. The other description for motion is made in terms of the spatial or current coordinates, called a spatial description or *Eulerian description.* The coordinate system is relative to a moving point in the body and hence is a *moving coordinate system.*

Motivated by the above discussions, let y be defined by the degradation velocity v as

$$\frac{dy}{dt} = v(t,y)$$

and assume

$$y(0) = \tilde{s}.$$

(That is, we assume that the observed intensity variable is initially equal to the 'true' FI in the absence of any FI decay.) For the remainder of the discussion presented here, $v(t,y)$ is taken to be the Gompertz

decay rate from Section 2.2.3. Thus we have the time-dependent change of coordinates

$$y = \tilde{s} - \frac{c}{k\log(10)}(1 - e^{-kt}).$$

Applying this change of coordinates to the left side of Equation (2.23),

$$\begin{aligned}
\frac{d}{dt}\left[10^{\tilde{s}}\log(10)n(t,\tilde{s})\right] &= \frac{d}{dt}\left[10^{y+\frac{c}{k\log(10)}(1-e^{-kt})}\log(10)n(t,y+\frac{c}{k\log(10)})\right] \\
&= \frac{d}{dt}\left[10^{y+\frac{c}{k\log(10)}(1-e^{-kt})}\log(10)\right]n(t,y+\frac{c}{k\log(10)}) \\
&\quad + \frac{dn}{dt}\left[10^{y+\frac{c}{k\log(10)}(1-e^{-kt})}\log(10)\right] \\
&= 10^{y+\frac{c}{k\log(10)}(1-e^{-kt})}\log(10)\frac{-ce^{-kt}}{n}(t,y+\frac{c}{k\log(10)}) \\
&\quad + \left(\frac{\partial n}{\partial t} - ce^{-kt}\frac{\partial n}{\partial y}\right)\left[10^{y+\frac{c}{k\log(10)}(1-e^{-kt})}\log(10)\right].
\end{aligned}$$

Plugging into Equation (2.23) and defining

$$\tilde{n}(t,y) = 10^{\left(y+\frac{c}{k\log(10)}(1-e^{-kt})\right)}\log(10)\cdot n\left(t, y+\frac{c}{k\log(10)}(1-e^{-kt})\right),$$

we have

$$\frac{\partial\tilde{n}}{\partial t} - ce^{-kt}\frac{\partial\tilde{n}}{\partial y} = -(\tilde{\alpha}(t,\tilde{s}) + \tilde{\beta}(t,\tilde{s}) - ce^{-kt})\tilde{n}(t,y) \tag{2.24}$$

$$+\chi_{(-\infty,\tilde{y}^*(t)]}2\tilde{\alpha}(t,\tilde{s}+\log_{10}2)\tilde{n}(t,\tilde{s}+\log_{10}2). \tag{2.25}$$

As before, we can eliminate the tildes for notational convenience. This demonstrates that the parameterization of the proliferation and death rates in terms of the 'translated coordinate' s is mathematically justified as a change of coordinates to a reference frame in which the measured FI of cells slowly decreases in time. It is worth noting that, from the y reference frame at time t, the coordinate $\tilde{s}$ is the point of intersection of the y-axis with the characteristic line passing through (t, y). The analysis can be made more general by assuming an unspecified change of variables $y = s + \nu(t, y)$. Provided the conditions of a global implicit function theorem [94] hold (so that it is possible to write $y = \mu(t, \tilde{s})$ for some function μ) then it is possible to obtain a general form of (2.25) [7].

The motivation for such a coordinate is the fact that, by accounting for the natural decrease of FI of the intracellular label, the quantity $s(t, y)$ is more strongly correlated with division number than the quantity y. Thus, the use of this 'translated coordinate' for the parameterization of α and β should provide a more accurate model of the observed data when compared to the simple implementation of spatial dependence. While this was found to be the case in [21] that analysis was done with an exponential label loss function. It remains to be shown that parameterization of the proliferation and death rates in terms of the translated coordinate s is advantageous when using the new model (2.21) and the Gompertz label loss function (2.19). To this end, four different parameterizations of the proliferation rate α and

two different parameterizations of the death rate function β are considered. In Section 2.4, results will be reported which demonstrate the effects of these different parameterizations on the effectiveness of the model.

First, we consider the simple case that $\alpha = \alpha(y)$. Given a fixed set of nodes $\{y_k\}$, we assume

$$\alpha = \alpha(y) = \sum_{k=1}^{K_\alpha} a_k l_k^{(\alpha)}(y), \tag{2.26}$$

where $l_k^\alpha(y)$ are piecewise linear spline functions satisfying

$$l_k^\alpha(y_j) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} .$$

It is assumed that $\alpha(0) = \alpha(3.5) = 0$. This assumption does not have a significant impact on the model as the nodes $\{y_k\}$ are chosen so that the proliferation rate can be varied as necessary at all values of the state variable where cells appear in the data. It does, however, add some measure of regularity to the computed proliferation rate function.

Alternatively, as discussed above, it may prove more accurate to use the translated coordinate s in order to represent the proliferation rate of cells with a particular division number. Given a fixed set of nodes $\{s_k\}$, we assume

$$\alpha = \alpha(s) = \alpha(s(t,y)) = \sum_{k=1}^{K_\alpha} a_k l_k^{(\alpha)}(s), \tag{2.27}$$

where the functions $l_k^\alpha(s)$ are defined as above. Again, it is assumed that $\alpha(0) = \alpha(3.5) = 0$.

We also consider the possibility that the proliferation rate depends explicitly on time. Indeed, we see in the data in Figure 1.1 that there is no proliferation at least during the first 24 hours of the assay. However, by $t = 48$ hours, it is clear that the population has begun to divide. Thus the assumption of time dependence seems appropriate. As above, we can still consider the proliferation rate either in terms of the state variable y or in terms of the translated coordinate s, in addition to its dependence on time. Given a set of nodes $\{y_k\}$ as above and a set of time nodes $\{t_m\}$, we parameterize

$$\alpha = \alpha(t,y) = \sum_{k=1}^{K_\alpha} \sum_{m=1}^{M} a_{km} l_k^{(\alpha)}(y) l_m^{(t)}(t), \tag{2.28}$$

where we now assume that the splines $l_k^{(\alpha)}(y)$ and $l_m^{(t)}(t)$ are piecewise linear in their respective variables. Again, we ensure some regularity in the inverse problem by requiring $\alpha(t,0) = \alpha(t,3.5) = 0$. It is also assumed that $\alpha(t,y) = 0$ for all $t \leq 24$ hours.

Finally, we consider the case that α is parameterized in time as well as the translated coordinate s. Given nodes $\{t_m\}$ as above and nodes $\{s_k\}$ in the translated variable, the proliferation rate function is then

$$\alpha = \alpha(t,s) = \alpha(t,s(t,y)) = \sum_{k=1}^{K_\alpha} \sum_{m=1}^{M} a_{km} l_k^{(\alpha)}(s) l_m^{(t)}(t), \tag{2.29}$$

where we again assume the splines $l_k^{(\alpha)}(y)$ and $l_m^{(t)}(t)$ are piecewise linear. As before, it is assumed $\alpha(t,0) = \alpha(t,3.5) = 0$ and $\alpha(t,s) = 0$ for all $t \leq 24$ hours.

Results from [21, 75, 77] indicate that the death rate function need not be quite as complex as the proliferation rate function. After the first few generations, the death rate of cells seems to be roughly constant. There is little reason to suspect that the death rate function depends on time and we do not consider it here. As before, we consider using the state variable y and the translated coordinate s to parameterize the death rate function. Given nodes $\{y_k\}$ (which may be distinct from the nodes used in the estimation of the proliferation rate α), we have

$$\beta = \beta(y) = \sum_{k=1}^{K_\beta} b_k l_k^{(\beta)}(y). \tag{2.30}$$

We assume $\beta(y) = b_1$ for all $y \in [0, y_1]$ and $\beta(y) = b_{K_\beta}$ for all $y \in [y_{K_\beta}, 3.5]$. Alternatively, using the translated coordinate, we have nodes $\{s_k\}$ and

$$\beta = \beta(s) = \beta(s(t,y)) = \sum_{k=1}^{K_\beta} b_k l_k^{(\beta)}(s) \tag{2.31}$$

with the assumptions $\beta(s) = b_1$ for all $s \in [0, s_1]$ and $\beta(s) = b_{K_\beta}$ for all $s \in [s_{K_\beta}, 3.5]$.

2.3 Parameter Estimation Procedure

Given the new model (2.21) and an appropriate parameterization of the proliferation and death rates, it is now possible to establish a framework for estimating the parameters of the model from a data set. However, as discussed in Section 1.2, the cytometry histograms show the numbers of cells counted into a given set of bins corresponding to particular ranges of log CFSE FI values in the $z = \log_{10} x$ coordinate while the model is a structured density in the variable $y = \log_{10}(x - x_a)$. Thus it is first necessary to determine a reasonable approximation of the smooth initial condition function $\tilde{\Phi}(y)$ using data from Day 0 ($t = 0$ hours). Next, an observation operator must be constructed which relates the model solution to the remaining histogram data ($t = 24$, 48, 96, and 120 hours).

2.3.1 Initial Condition Construction

Recall from Chapter 1 that, at time t_j, the data is stored as a set of ordered pairs (z_k^j, n_k^j), $k = 1, \ldots, K(j)$ which correspond to the number of cells n_k^j counted into the bin with left boundary z_k^j. To construct the initial condition, a smooth line is drawn through the data at $t = 0$, and ordered pairs representing the line can then be determined using DataThief [101]. These smoothed histogram curves are scaled upward into a smooth initial condition density $\hat{\Phi}(z)$ so that the total label content is the same for the smooth density as for the original histogram data. The results are depicted in Figure 2.6. Finally, given the initial condition $\hat{\Phi}(z)$, the initial condition function for $\tilde{n}(t,y)$ is computed by noting that

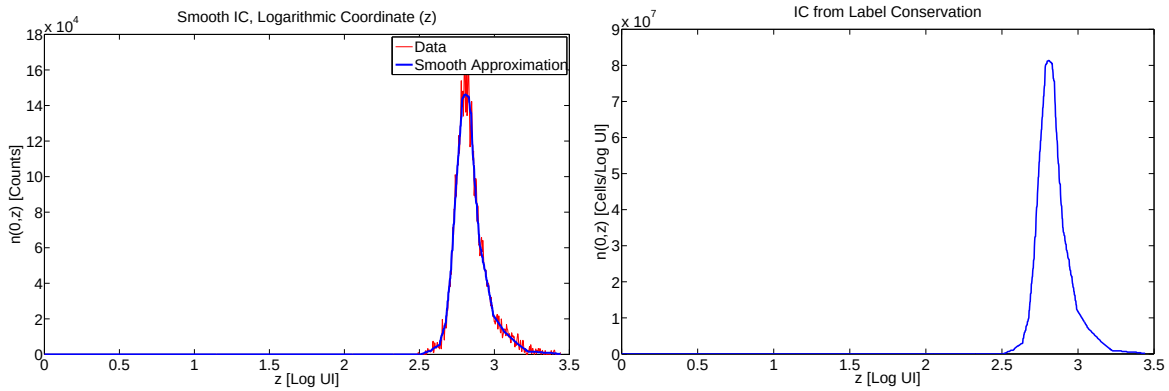

Figure 2.6: Left: Smoothed histogram data at $t = 0$ in the z coordinate. Right: Density function $\hat{\Phi}(z)$ computed from the smoothed histogram data.

$y = \log_{10}(10^z - x_a)$ and using the label-preserving identity

$$\hat{\Phi}(z) = \frac{10^z}{10^z - x_a}\tilde{\Phi}(y(z)). \tag{2.32}$$

This, then, provides an initial condition for (2.21).

2.3.2 Observation Operator

While the procedure above provides a means of determining a smooth initial condition density $\tilde{\Phi}(y)$ from the smoothed histogram data, the comparison of the model, a density defined in terms of y, to the data, a histogram in terms of z, represents the opposite problem. In order to make this comparison, we need to perform two steps. First, we must transform the structured density $\tilde{n}(t, y)$ into a function of z. This is done analogously to the transformation (2.32) above. Second, we must transform this structured density into histogram numbers. To do this, we note that at time t_j, the total number of cells with FI between z_k^j and z_k^{j+1} is

$$I[\hat{n}](t_j, z_k^j) \equiv \int_{z_k^j}^{z_k^{j+1}} \hat{n}(t_j, z)dz \approx \left[\frac{\hat{n}(t_j, z_k^{j+1}) + \hat{n}(t_j, z_k^{j+1})}{2}\right]\left(z_k^{j+1} - z_k^j\right), \tag{2.33}$$

where the trapezoid rule has been used to approximate the integral. In general, we find that this is an effective method of obtaining histogram counts from the smooth density model solution. Implications of the trapezoidal approximation on the inverse problem are considered at greater length in Chapter 3. However, the varying sizes of the bins used to record the available data set poses somewhat of a problem. While the bins are generally regularly spaced, there are a few bins randomly placed in the data which are either much larger or much smaller than the neighboring bins. As a result, the histogram data $I[\hat{n}](t_i, z_j)$ computed from the smooth densities exhibits large jumps up or down at these points. This is strictly the result of the irregular bin sizes present in the data set and not the model solution

itself. These jumps are problematic for the OLS procedure discussed below and as such these bins are removed from the data set. We emphasize again that, in future data sets, the bins can be set as needed, rather than being fixed in advance.

2.3.3 Numerical Method

While (2.21) is defined on an infinite domain, all cells in the population maintain FI sufficiently greater than x_a, so that it is acceptable to solve (2.21) only on the domain $y \in [0, y_{\max}]$. In practice, we set $y_{\max}$ independent of the parameter x_a and thus solve the equation (2.21) on the same computational domain regardless of the parameters (i.e., those passed in by the nonlinear optimization solver). Once the solution $n(t, y)$ is computed, it is possible to use (2.32) again to change variables back to $\hat{n}(t, z)$ for comparison to the data.

For the current data set, we use 512 evenly spaced nodes in the interval $y \in [0, 3.5]$. The forward solution is computed using a publicly available hyperbolic PDE solver written by L. Shampine which implements the Lax-Wendroff scheme.

2.3.4 Ordinary Least Squares Framework

Given the appropriate parameterizations of α and β, we now have a complete set of parameters $\theta = (x_a, c, k, \{a\}, \{b\})$ which define the model solutions. Thus in the analysis below we think of the parameterized model $\tilde{n}(t, y; \theta)$ satisfying (2.21) given parameter θ. We now turn our attention to an inverse problem procedure which seeks to determine the parameters best describing the available data.

Following standard inverse problem procedure for ordinary least squares (OLS) [22, 35, 37], we assume that the data n_j^k represent an observation of the model solution evaluated at the true parameter θ_0 with the addition of some amount of noise. Thus, we can consider the data as a random variable

$$N_k^j = I[\hat{n}](t_j, z_k^j; \theta_0) + \mathcal{E}_{kj}, \tag{2.34}$$

where $\{\mathcal{E}_{kj}\}$ are random variables with $E[\mathcal{E}_{kj}] = 0$ and $Var(\mathcal{E}_{kj}) = \sigma^2$. We remark that the assumption of constant variance for the error terms is standard for OLS formulations of inverse problems. One can examine the accuracy of such an assumption ex post facto through the use of residual-based statistical tests [8, 22, 96]. In [21], such an analysis revealed that the actual error variance was neither constant nor proportional to the square of the model solution (a 'relative error model'). For the moment, we remark that the OLS assumption of constant variance, while possibly not exactly correct and hence not adequate for use in asymptotic parameter distributional analysis, is sufficient to provide a basis for computational parameter estimation, which will demonstrate the ability of the current model to fit the available data set. Given our uncertainty regarding the exact nature of the observed error process, we postpone a more detailed analysis of uncertainty in the estimated parameters (standard errors, confidence intervals, etc.) [8] for future work. Further analysis and characterization of the experimental measurement error with implications for the least squares framework can be found in Chapters 4 and 5.

Table 2.3: Summary of parameters $\theta = (x_a, c, k, \{a\}, \{b\})$ which define the model solution, with minimum values, maximum values, and units. Forward simulations of the model demonstrate the reasonableness of the bounds provided.

Parameter	Minimum	Maximum	Units
a_i	0	1	hr^{-1}
b_i	0	1	hr^{-1}
x_a	0	100	UI
c	0	0.1	UI/hr
k	0	0.005	hr^{-1}

Given the statistical model (2.34), we can write the data as realizations

$$n_k^j = I[\hat{n}](t_j, z_k^j; \theta_0) + \epsilon_{kj} \tag{2.35}$$

of the random variables (2.34). The goal of the OLS procedure is the determination of the parameter θ which minimizes the sum of squared residuals. Given the random variables N_k^j from (2.34), the OLS estimator is

$$\theta_{\text{OLS}} = \arg\min_{\theta \in \Theta} \sum_{j=1}^{J} \sum_{k=1}^{K(j)} (I[\hat{n}](t_j, z_k^j; \theta) - N_j^k)^2 = \arg\min J(\theta), \tag{2.36}$$

where Θ is a set of admissible parameters for the model (see Table 2.3). As the data n_k^j are realizations of the random variables N_k^j, it follows that the OLS estimate

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\theta \in \Theta} \sum_{j=1}^{J} \sum_{k=1}^{K(j)} (I[\hat{n}](t_j, z_k^j; \theta) - n_k^j)^2 = \arg\min J(\theta), \tag{2.37}$$

is a realization of the OLS estimator θ_{OLS}. This optimization was carried out with the MATLAB constrained optimization routine `fmincon`, which implements the BFGS algorithm at each step to solve a quadratic subproblem. Because such routines can become trapped in local minima, several initial iterates were tried for each optimization.

2.4 Results

The primary uncertainty in the inverse problem procedure is the choice of nodes $\{y_k\}$ or $\{s_k\}$ for the estimation of the proliferation and death rates. We have no a priori information as to how many nodes should be used nor any information regarding where those nodes should be placed. To illustrate this point, in Figure 2.7 we depict the estimated proliferation and death rate functions given three different choices of nodes $\{y_k\}$. First, seven nodes were evenly spaced in the interval $[1.125, 2.925]$. Next the number of nodes was increased to 13 so that the separation between nodes was halved (and so that the increase in parameters is a refinement to the model). This procedure was repeated and the estimation was performed a third time with 25 nodes. Intuitively, each refinement must provide a more accurate fit of the model to the data, as the total OLS cost cannot increase as the number of parameters is increased.

Table 2.4: Average proliferation rates (in units 1/hr) in terms of numbers of divisions undergone, computed from Figure 2.7. Using Figure 1.1, approximate ranges (in the coordinate z) corresponding to particular division numbers are determined. These are then used to compute the corresponding ranges in the variable y using the estimated level of cellular autofluorescence. In spite of the differences in the numbers of nodes used in the parameterization of the proliferation rate function $\alpha(y)$, average proliferation values are estimated consistently for each generation.

Division Number	z-axis Range	7 Nodes	13 Nodes	25 Nodes
6	$[0.00, 1.05]$	0.0016	0.0015	0.0016
5	$[1.05, 1.30]$	0.0043	0.0050	0.0052
4	$[1.30, 1.60]$	0.0094	0.0103	0.0108
3	$[1.60, 1.90]$	0.0166	0.0174	0.0206
2	$[1.90, 2.25]$	0.0325	0.0308	0.0314
1	$[2.25, 2.55]$	0.0284	0.0198	0.0266
0	$[2.55, 3.50]$	0.0047	0.0097	0.0072

However, we also see in Figure 2.7 that as the number of nodes increases, the estimated proliferation rate function becomes less regular. Conversely, we see that some measure of regularity can be imposed on the function $\alpha(y)$ by choosing the proper set of nodes with which to estimate it (a so-called 'regularization by discretization' [15, 16]). While we do have available residual-sum-of-squares based statistical tests [8, 10, 22] to quantify the improvement in the model with each refinement, we choose to balance this additional information with a desire to estimate a semi-regular function α.

It is worth remarking further that the increasingly complex structure of the function $\alpha(y)$ (as the number of nodes is increased) is only a relic of the estimation procedure and has nothing to do with any meaningful information regarding the population of cells being studied. In order to verify this, in Table 2.4 we present, for each parameterization of the function $\alpha(y)$ discussed in the previous paragraph, the average value of the estimated proliferation rate in terms of the numbers of divisions the cells have undergone. To compute these values, Figure 1.1 is used to determine approximate ranges (in the coordinate z) corresponding to each generation of cells. By changing these ranges from the variable z to y (in which the estimation of the proliferation rate function was performed), the average value of the proliferation rate function can be determined in each range. As seen in Table 2.4, the estimates are reasonably consistent regardless of the number of nodes used in the estimation.

Given the above discussion, we choose 13 nodes for the estimation of the proliferation rate function $\alpha(y)$ and 5 nodes for the estimation of the death rate function $\beta(y)$ in an effort maximize the flexibility of the estimation while also maintaining some regularity in the estimated functions. The OLS best-fit solution is shown in Figure 2.8. The optimal proliferation and death rate functions are shown graphically in the center panels of Figure 2.7, with numerical values provided in Tables 2.5 and 2.6. The total OLS cost for the estimation is $J(\hat{\theta}_{\text{OLS}}) = 1.7270 \times 10^{12}$, with $x_a = 8.2316$, $c = 5.6169 \times 10^{-3}$, and $k = 1.1203 \times 10^{-8}$.

We observe that, while the OLS best-fit solution for time-independent proliferation is accurate for $t = 96$ and 120 hours, the model predicts far too many cells with high generation number at $t = 24$ and 48 hours. This seems to be a manifestation of the absence of a delay (in the form of time dependence) between the time cells are stimulated and the time at which those stimulated cells divide. Thus, in

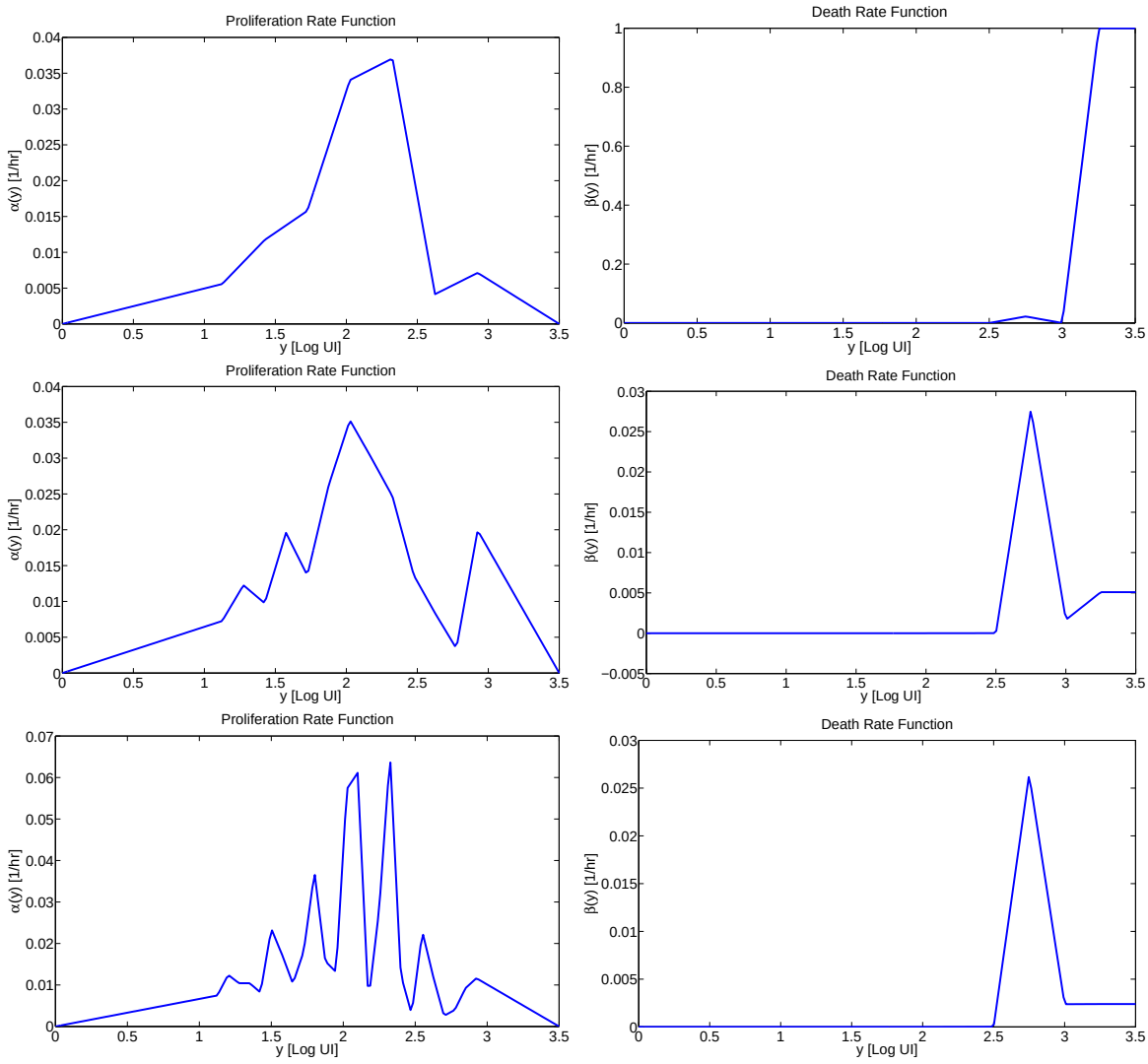

Figure 2.7: Left: Estimated proliferation rate function $\alpha(y)$ for three difference choices of nodes. Top: 7 nodes evenly spaced in [1.125,2.925]. Middle: 13 nodes evenly spaced in [1.125,2.925]. Bottom: 25 nodes evenly spaced in [1.125,2.925]. Note that, while the overall shape of $\alpha(y)$ remains largely the same, the middle figure seems to provide the most information while remaining some semblance of regularity. These functions can be used to determine the average rate of proliferation in terms of the number of divisions undergone (Table 2.4). Right: the corresponding estimated death rate function $\beta(y)$, estimated using 5 fixed nodes in each case.

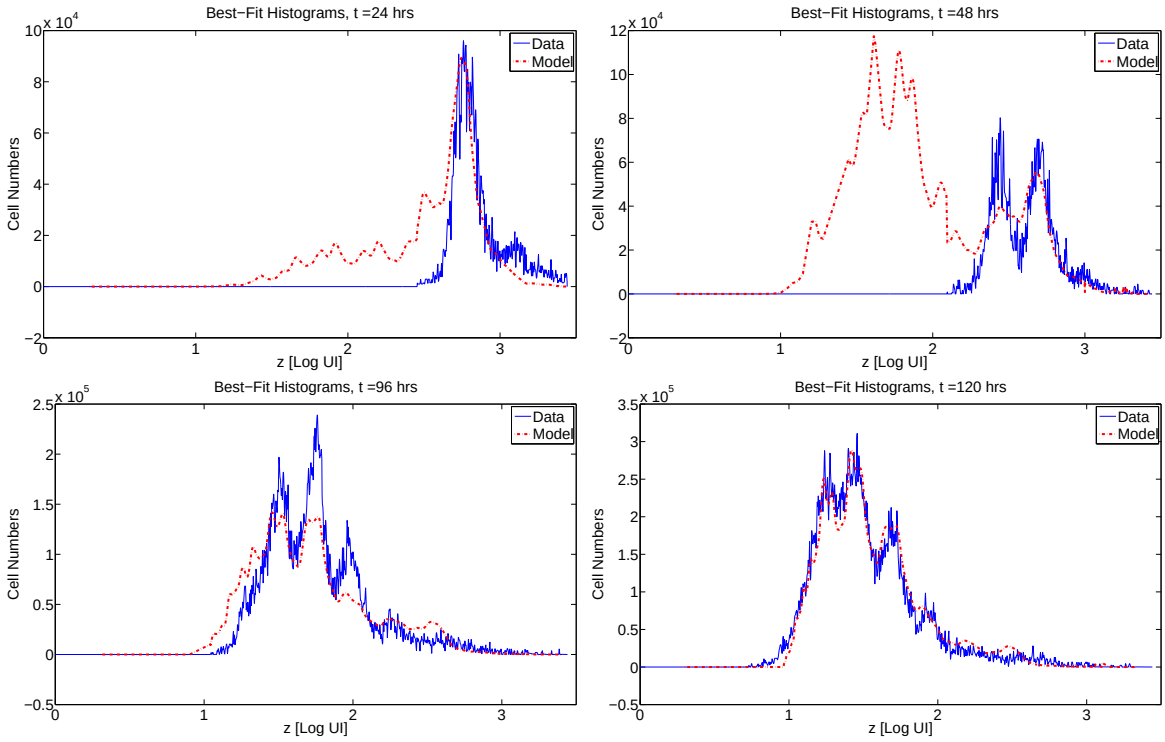

Figure 2.8: OLS best-fit solution with $\alpha = \alpha(y)$ (13 nodes), $\beta = \beta(y)$ (5 nodes). 21 total parameters in the model, total cost $J(\hat{\theta}_{\text{OLS}}) = 1.7270 \times 10^{12}$. While the model clearly is not accurate in allowing far too many cells with large division number too early in time, the *locations* of the division peaks along the horizontal axis are quite accurate, in support of the role of autofluorescence as well as the Gompertz decay of label.

Table 2.5: Results for the OLS estimation of $\alpha(y)$ with 13 nodes. This estimated proliferation rate function is shown graphically in the left-center panel of Figure 2.7. Average rate of division in terms of division number is computed in Table 2.4.

$y_k^{(\alpha)}$	a_k
1.1250	0.0073
1.2750	0.0123
1.4250	0.0097
1.5750	0.0196
1.7250	0.0136
1.8750	0.0262
2.0250	0.0353
2.1750	0.0301
2.3250	0.0247
2.4750	0.0137
2.6250	0.0084
2.7750	0.0034
2.9250	0.0199

Table 2.6: Results for the OLS estimation of $\beta(y)$ with 5 nodes when $\alpha = \alpha(y)$ is estimated with 13 nodes. This estimated death rate function is shown graphically in the right-center panel of Figure 2.7.

$y_k^{(\beta)}$	b_k
2.0000	0.0000
2.5000	0.0000
2.7500	0.0278
3.0000	0.0017
3.2500	0.0051

addition to the arguments of Section 2.2.5, we have a mathematical rationale for the incorporation of time dependence into the proliferation rate. While the model does not accurately count the numbers of cells in the earlier time points, we do remark that the Gompertz label loss model and the incorporation of cellular AutoFI do accurately predict the *location* (along the horizontal axis) of the subsequent generations of cells in culture. Thus, it appears safe to conclude that the parameter γ from [21, 77] has been effectively removed from any modeling needs.

Given this discussion, we next consider the possibility that the proliferation rate function $\alpha(t, y)$ may depend on time as well as on the structure variable y (see Section 2.2.5) in order to better estimate the numbers of cells in each generation at a given time. We would like to make use of model refinement techniques in order to quantify the resulting improvement in the fit of the model to data while also accounting for the increased complexity of the model. Thus, we use the same 13 nodes as above for the structural discretization of α. For the time discretization, nodes $\{t_m\} = [48, 60, 72, 96, 120]$ are used. The death rate function $\beta(y)$ is estimated exactly as before. This parameterization results in a model with 73 parameters. After calibration to the data, the resulting cost is $J(\hat{\theta}_{\text{OLS}}) = 3.1302 \times 10^{11}$ with $x_a = 6.2698$, $c = 4.8123 \times 10^{-3}$, and $k = 9.8091 \times 10^{-8}$; the fit of the model to the data is shown in Figure 2.9. It is clear from the figure that the improvement in fitting the model to the data is quite significant. Moreover, because the inclusion of time-dependence is a refinement of the time-independent model, residual-sum-of-squares-based statistical tests exist to quantify whether the increase in complexity of the model (from 21 to 73 parameters) is justified by the resulting reduction in cost. Using the method described in [22, Ch. 3], we find that the time-independent model can be rejected in favor of time-dependent proliferation with very high ($> 99.999\%$) confidence.

Thus we see that, once the proliferation rate is allowed to vary as a function of time, the model very closely mimics the data in terms of the numbers of cells in each generation at a given time. Moreover, we again point out that the physiological explanation for the dilution of FI by division, as well as the Gompertz model for natural FI decay do an excellent job of predicting the locations (along the horizontal axis) of each generation of cells. Still, we continue further to consider one more potential improvement to the model. Following the analysis of [21] and the discussion of Section 2.2.5, we consider parameterizing the proliferation and death rate functions in terms of the 'translated coordinate' s. As discussed previously, it is expected that this coordinate correlates much more closely with division number than the coordinates z or y. As such, estimation of the proliferation and death rates in terms of this quantity should provide a more meaningful (and less biased) estimate when these functions are analyzed in terms

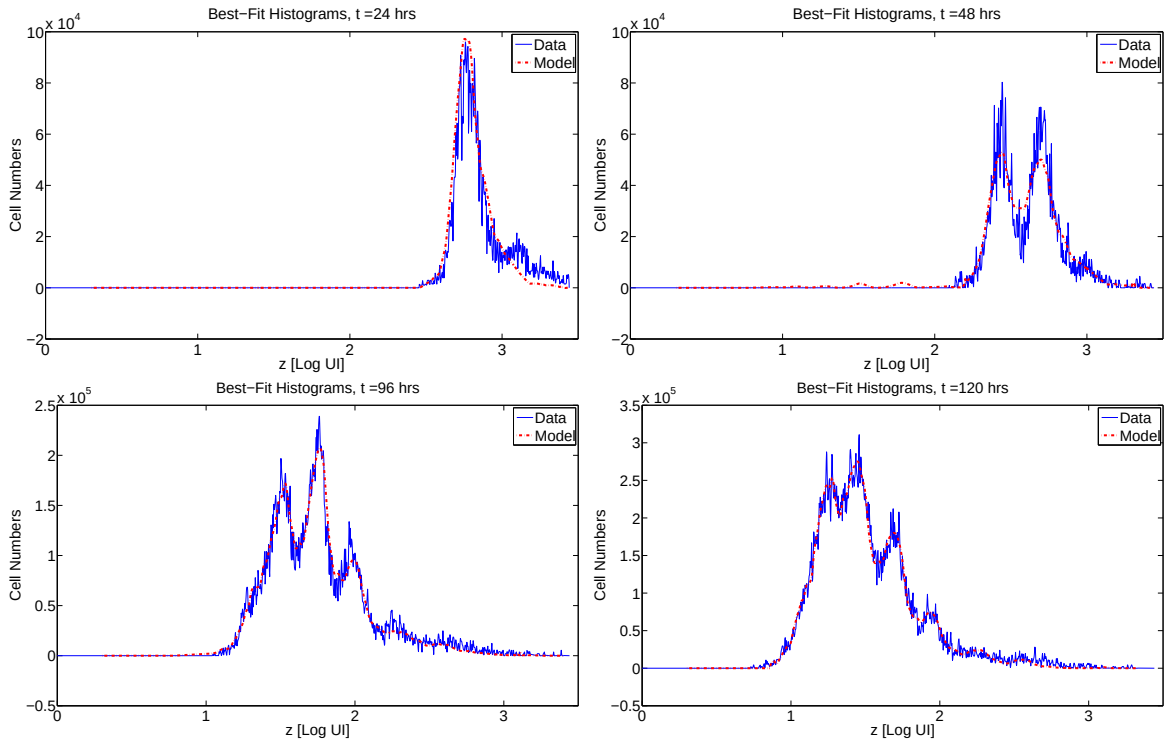

Figure 2.9: OLS best-fit solution with $\alpha = \alpha(t, y)$, $\beta = \beta(y)$. 73 total parameters in the model, total cost $J(\hat{\theta}_{\text{OLS}}) = 3.1302 \times 10^{11}$.

of division number (in the manner of Table 2.4). It is worth noting that the parameterization of the functions α and β in terms of s is not a model refinement (compared to parameterization in terms of y) so that simple model refinement-based statistical tests are not applicable. While additional (e.g. information-theoretic) tests could be used, we forgo that analysis here in the interest of brevity. However, as will be shown, parameterization in terms of s does in fact provide a more meaningful correlation between the estimated cell turnover rates and division number (in additional to providing a slightly lower cost!), which justifies its use.

The nodes used from the proliferation and death rate functions, as well as the estimated rates at those nodes, are given in Tables 2.7 and 2.8, respectively, and the functions are shown graphically in Figures 2.10 and 2.11. As before, 73 parameters arise in this parameterization of the model. The total cost is $J(\hat{\theta}_{\text{OLS}}) = 3.0901 \times 10^{11}$, with $x_a = 6.4053$, $c = 5.5246 \times 10^{-3}$, and $k = 5.0323 \times 10^{-4}$; the fit of the model to the data is shown in Figure 2.12.

Visually, the fit of this model (using s for the structure discretization of the proliferation and death rates) is comparable to the previous model (using y), and the cost is slightly lower. The significant advantage in using s, as noted above, is that the translated coordinate s is more strongly correlated with division number. To see this, the data from Figure 1.1 are shown in the translated coordinate in Figure 2.13. While there is still some overlap among the generations of cells in the histogram data,

Table 2.7: Results for the OLS estimation of $\alpha(t, s)$. This estimated proliferation rate function is shown graphically in Figure 2.10. Average rate of division in terms of division number is computed in Table 2.14.

	t_k				
$s_k^{(\alpha)}$	48	60	72	96	120
1.1875	0.0713	0.1404	0.0000	0.0000	0.0167
1.3375	0.2028	0.0522	0.0001	0.0000	0.0175
1.4875	0.6036	0.0303	0.0376	0.0009	0.0281
1.6375	0.2896	0.0138	0.0004	0.0075	0.0251
1.7875	0.0618	0.0001	0.0000	0.0409	0.0220
1.9375	0.0091	0.0345	0.0020	0.0119	0.0220
2.0875	0.0837	0.0002	0.0400	0.0326	0.0391
2.2375	0.0018	0.1956	0.0083	0.0001	0.0231
2.3875	0.0050	0.0059	0.0962	0.0394	0.0463
2.5375	0.0000	0.1949	0.0128	0.0050	0.0000
2.6875	0.0155	0.1101	0.1528	0.0422	0.0239
2.8375	0.0351	0.0446	0.0000	0.0000	0.0000
2.9875	0.0346	0.0000	0.0012	0.1317	0.0092

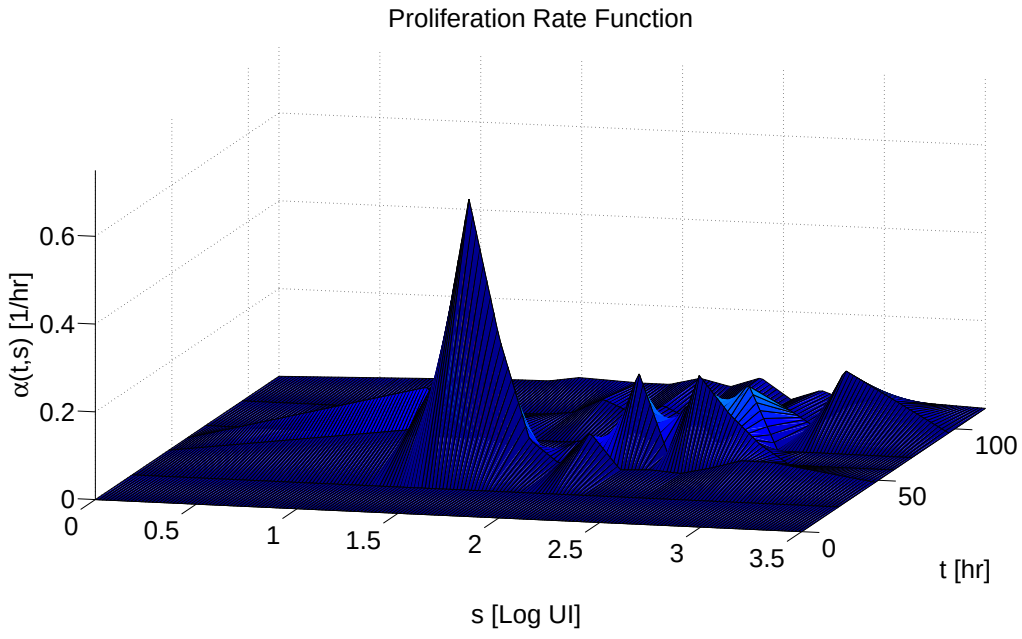

Figure 2.10: OLS best-fit proliferation rate function $\alpha(s, t)$.

Table 2.8: Results for the OLS estimation of $\beta(s)$ when $\alpha = \alpha(t, s)$. This estimated death rate function is shown graphically in Figure 2.11

$s_k^{(\beta)}$	b_k
2.0000	0.0054
2.5000	0.0000
2.7500	0.0238
3.0000	0.0061
3.2500	0.0000

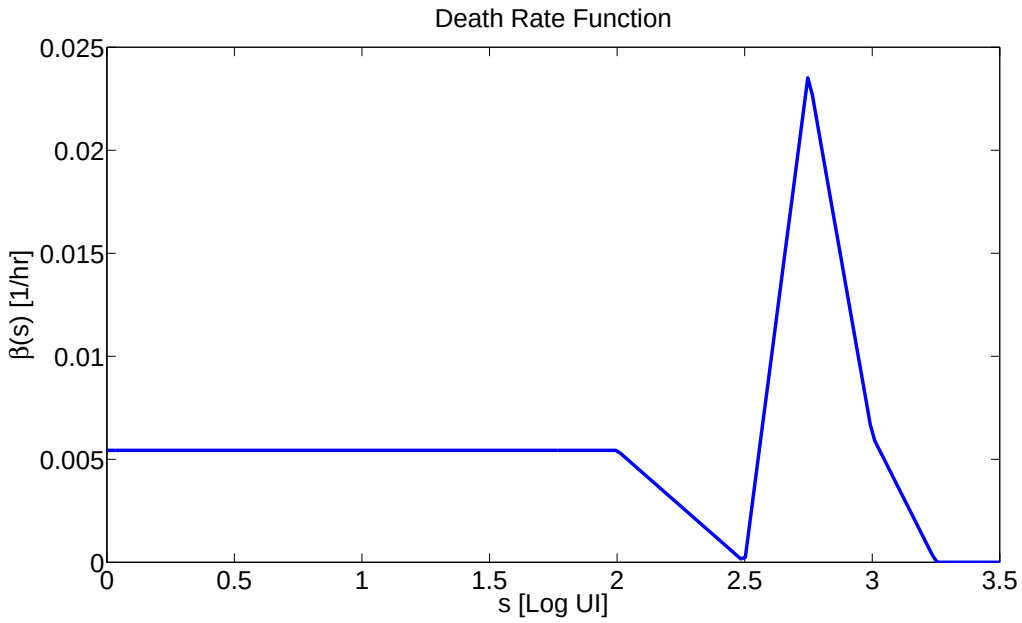

Figure 2.11: OLS best-fit death rate function $\beta(s)$ when the proliferation rate is assumed to depend on both s and t.

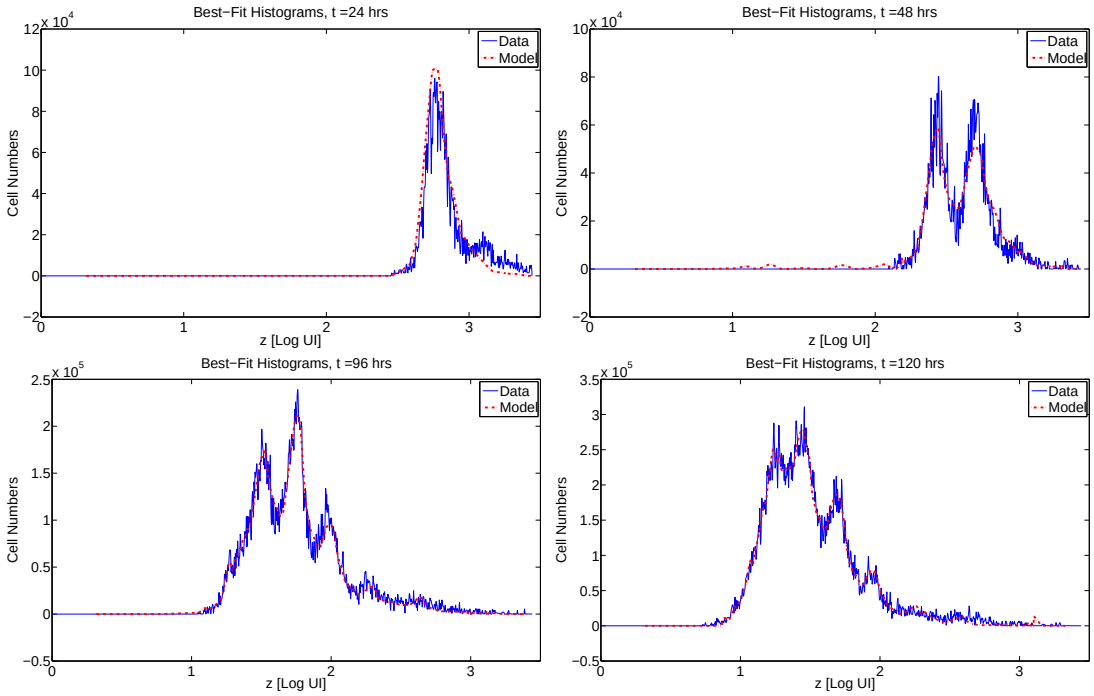

Figure 2.12: OLS best-fit solution with $\alpha = \alpha(t, s)$, $\beta = \beta(s)$. 73 total parameters in the model, total cost $J(\hat{\theta}_{\mathrm{OLS}}) = 3.0901 \times 10^{11}$.

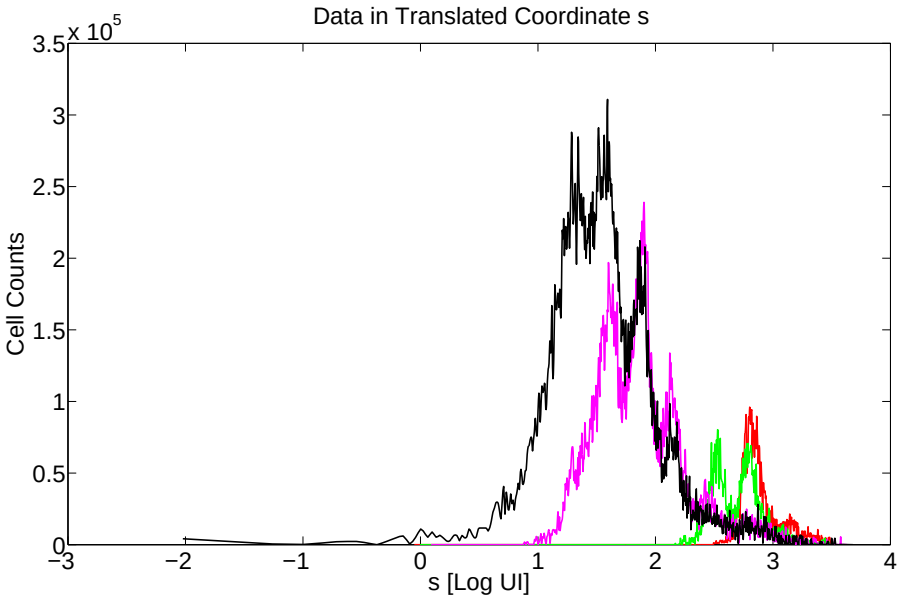

Figure 2.13: Data for $t = 24, 48, 96, 120$, respectively, plotted in the translated coordinate s. Observe that the peaks corresponding to distinct division numbers closely align.

the translated coordinate provides an axis on which cells do not drift as they slowly lose CFSE FI. Particularly when compared to Figure 1.1, we see that it is much easier to assign distinct regions of the s axis to particular division numbers when compared to using the z (and thus y) axis. Moreover, because cells do not drift to the left on the s axis, regions assigned to particular division numbers remain valid for all time.

Given the near-alignment of the generations of cells in the translated coordinate s, a similar analysis to that presented in Table 2.4 can be performed. By determining intervals (in the s coordinate) corresponding to particular division numbers, the average rate of proliferation for cells having undergone a specified number of divisions can be computed. Because we now have a proliferation rate which depends explicitly on time, we compute the average proliferation rate (in terms of the number of divisions undergone) and display this information as a function of time (rather than averaging in time as well) in Figure 2.14, thus preserving what we believe to be an important feature of the population of cells (that the proliferation rates change in time). When estimating the time-dependent proliferation rate, it should be noted that, for high division number, the rate estimated for early times must be interpreted with caution: the rate is ultimately meaningless until cells have emerged in the population which divide at that rate. One potential solution to this caveat is to use a more complex (e.g. non-rectangular) grid for the estimation of $\alpha(t, s)$.

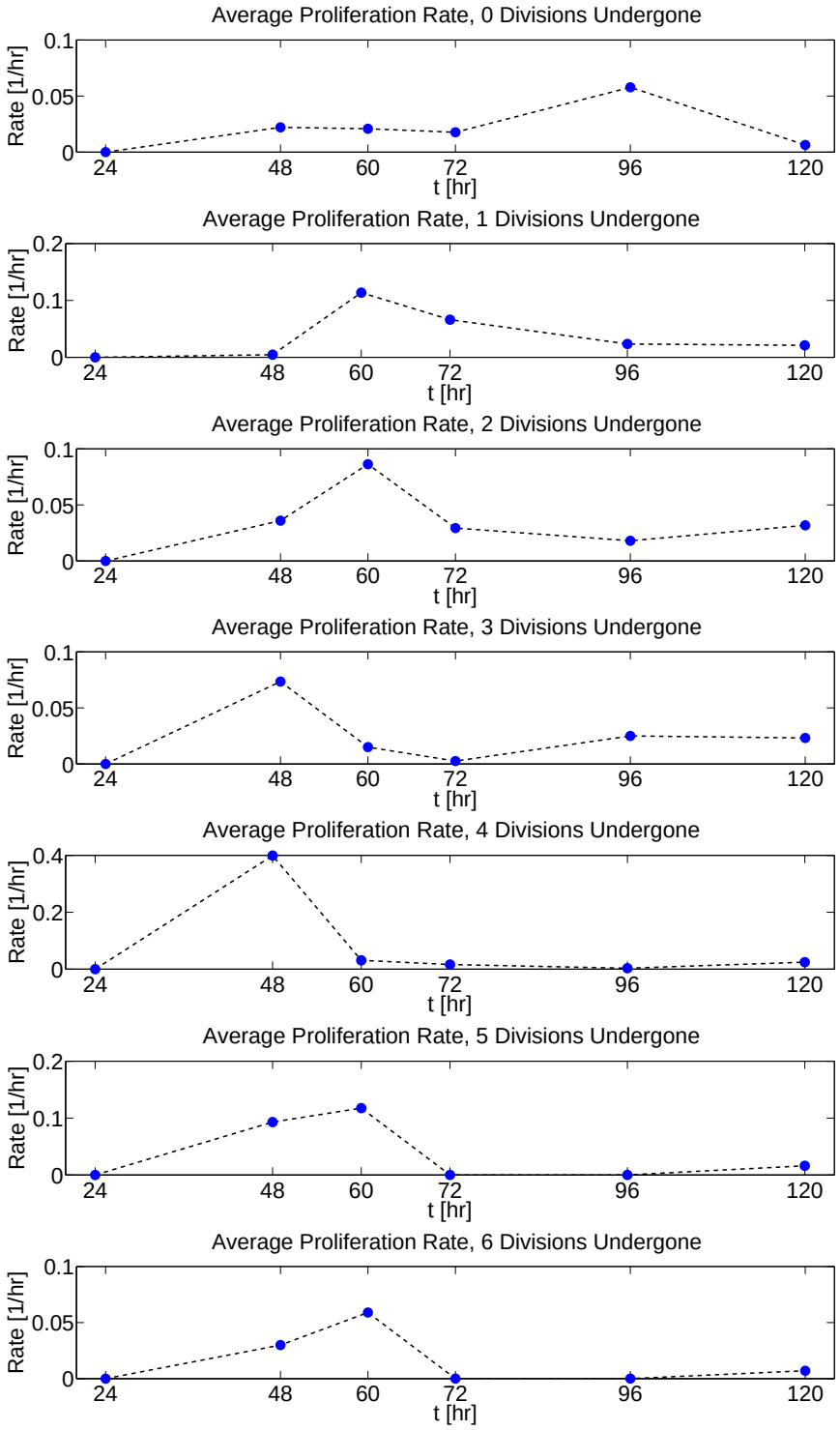

Figure 2.14: Average proliferation rate as a function of time for each generation of cells.

2.5 Discussion and Remarks

In this chapter, we began with the label-structured PDE model (2.2) from [21] and [77] and have presented significant modifications and clarifications. Of primary importance is that the parameter γ, used previously to heuristically explain the dilution of CFSE resulting from cell division, has been replaced by a physiologically-based mechanism which accounts for cellular autofluorescence. Also important is the use of the Gompertz decay process to explain the natural decay of CFSE observed in the data. We have seen that these two improvements in the model are fully capable of fitting a CFSE data set. Using the data set from [21, 77], we have shown that the incorporation of autofluorescence and Gompertz decay of label provide a mathematical model with firm physiological underpinnings which can accurately describe CFSE histogram data directly. Moreover, these revisions provide clarity to the model because they can be understood in terms of physiologically relevant and easily observable features of the data.

In support of the results of [21], parameterization of the proliferation and death rates α and β in terms of the translated or moving variable s provides an improvement to the OLS fit of the model to the data. Beyond this minor improvement, the introduction of this variable was motivated by the fact that, when all data is placed in the coordinate s (Figure 2.13), the peaks corresponding to distinct division numbers align much more closely than when presented in terms of the measurement variable z (Figure 1.1). As such, the estimation of the proliferation and death rate functions more directly relates the dependence of division and death rates on division number. While the exact shape of the estimated proliferation and death rate functions may change with various choices of nodes (Figure 2.7), we have seen that, for reasonable parameterization, the average proliferation rates (for each division number) are consistently estimated (Table 2.4) and the dependence of these rates on time can be explored (Figure 2.14). Thus, it is believed that the translated coordinate s permits the estimated functions α and β to be easily related to intuitive measures of a lymphocyte response such as mean division time, mean doubling time, etc. Such information provides a nearly complete quantitative picture of a dynamic T cell responsiveness and thereby may be helpful for mechanistic studies of immune control. Because of the nonparametric manner in which the proliferation and death rates are estimated, this model is able to encapsulate a wide variety of proliferative responses as various types of cells are subjected to a variety of experimental conditions and then measured.

The parameters of the revised model (2.21) are estimated in an ordinary least squares framework from a time-series of CFSE histograms. Because the data used in this report was already in histogram form when it was received, the irregular size of certain bins caused some computational difficulty. This problem should not be present in future data sets, when the histogram bin spacing can be directly controlled to remove these difficulties. As the ultimate goal of any model of the immune response is the comparison of changing intra- and extracellular conditions on proliferative behavior [51, 55], uncertainty quantification in the form of confidence intervals are necessary to facilitate such a comparison. Asymptotic theory for sum-of-squares-based estimators and model comparison tests [8, 10, 35, 96] exist, but rely upon a correct underlying statistical model for the data. As acknowledged in Section 2.3.4, the use of an ordinary least squares framework is premised upon the assumption that the 'noise' in the data has constant variance. A detailed analysis of the residuals is postponed until after the development of a compartmental model in Chapter 3, and implications for the quantification of uncertainty in the parameter estimates are

discussed in Chapters 4 and 5. It should be noted that while the determination of an accurate statistical model is of vital importance for the unbiased estimation of standard errors and confidence intervals for the estimated model parameters [22], a slight misspecification of the error model does not invalidate the ability of the model to accurately fit the data.

In addressing these issues, particular attention will need to be paid to the mechanism responsible for the apparently spurious measurement at $t = 72$ hours. As was shown in Section 1.2, the total quantity of CFSE FI in the cell culture is measured to increase from $t = 48$ hours to $t = 72$ hours, a physical impossibility. For that reason, the data collected at $t = 72$ hours is completely omitted from the current analysis. The inability of the current model to describe this behavior is not directly a shortcoming of the model itself (as any method, e.g., deconvolution with a series of gaussian curves, would suffer from a similarly biased result) but rather represents an incomplete understanding of the nature of the observation process. Indeed, it is an asset of the current model, derived from conservation principles, that such a feature has even been noticed. In Chapter 4, it is shown that variability in the measurement procedure (specifically, in the numbers of counted beads used to scale the sampled population) is most likely responsible for the apparent violation of the conservation law. In Chapter 5, a new statistical model/observation process is considered which might resolve this discrepancy.

The mathematical model presented here is more complex than many of the existing frameworks for understanding cell turnover kinetics. While the number of parameters will vary depending upon the manner in which nodes for the estimation of the proliferation and death rate functions α and β are chosen, the best-fit results presented in this report were obtained with 73 parameters. Optimization times range from 1 to 8 hours, depending upon the accuracy of the initial iterate and the error tolerances selected for the optimization routine. In spite of this additional complexity, the current model accurately predicts CFSE-based proliferation dynamics and does so by directly addressing histogram data from the assay. By avoiding the need for any deconvolution techniques to extract cell numbers (per generation) from the histograms, some potential bias and/or error is avoided. Additionally, by directly addressing quantities such as autofluorescence and the natural decay of label, their effects on the observed behavior of the population can be quantified.

The one major limitation of the new model is its inability to accurately determine the numbers of cells having undergone a specified number of divisions at a given time. In general, the number of cells having measured FI in a given range $[z_1, z_2]$ is

$$\int_{z_1}^{z_2} n(t,x)dx$$

where $n(t,x)$ is determined from the model solution $\tilde{n}(t,y)$ using the inverse of the change of coordinates discussed in Section 2.2.4. However, it is clear Figure 2.13 that consecutive generations of cells in the data (adjacent peaks in the histograms) share too much overlap for such a computation to be accurate. One could use a more complex deconvolution technique, but it was exactly such techniques that the current efforts were undertaken to avoid. In the next chapter, it is shown that a simple reformulation of the current PDE model can be used to obtain a compartmental model which not only fits the data at least as well as the current PDE model, but also can be used to obtain accurate cell counts.

Chapter 3

A Compartmental Model to Compute Numbers of Cells

3.1 Motivation

In the previous chapter, results from [20, 21, 75, 77] were revised and extended to form a physically and biologically motivated structured partial differential equation population model which could be fit accurately and directly to flow cytometry data. As discussed there, this is a significant step in establishing a mathematical framework for the quantitative description of an immune response which does not rely upon a deconvolution of the data into numbers of cells per generation (contra [38, 39, 40, 50, 51]). The revised PDE model is a fragmentation equation which relates the structured population density $n(t,x)$ to the rates of proliferation $\alpha(t,x)$ and death $\beta(t,x)$,

$$\frac{\partial n(t,x)}{\partial t} - ce^{-kt}\frac{\partial[(x-x_a)n(t,x)]}{\partial x} = -(\alpha(t,x)+\beta(t,x))n(t,x)+\chi_{[x_a,x^*]}4\alpha(t,2x-x_a)n(t,2x-x_a), \quad (3.1)$$

where the structure variable x is the fluorescence intensity (in arbitrary units of intensity, UI) of the cells resulting from a quantity of CFSE within those cells. Thus we refer to this as a label structured population model (as opposed to age or size structure, etc. [85]). The loss of FI in time as a result of the natural decay of the intracellular proteins to which the fluorescent CFSE conjugates bind is approximated with a Gompertz decay process [69] with parameters c and k, which was shown to accurately describe the biphasic decay [84, 89, 104] of CFSE FI observed in data sets. The parameter x_a represents the natural autofluorescence intensity of cells in the absence of CFSE.

The goal of such a mathematical model is to provide biologists with simple yet intuitive and meaningful parameters with which a population of cells can be described. In particular, information such as average rates of division and cell viability are essential to the analysis of the effects of changing experimental conditions on proliferative behavior. The motivation for the use of FI as a structure variable is that the serial dilution of CFSE by cell division creates a correlation between measured FI and the number of divisions a cell has undergone. This motivating assumption is accurate to a degree, as one

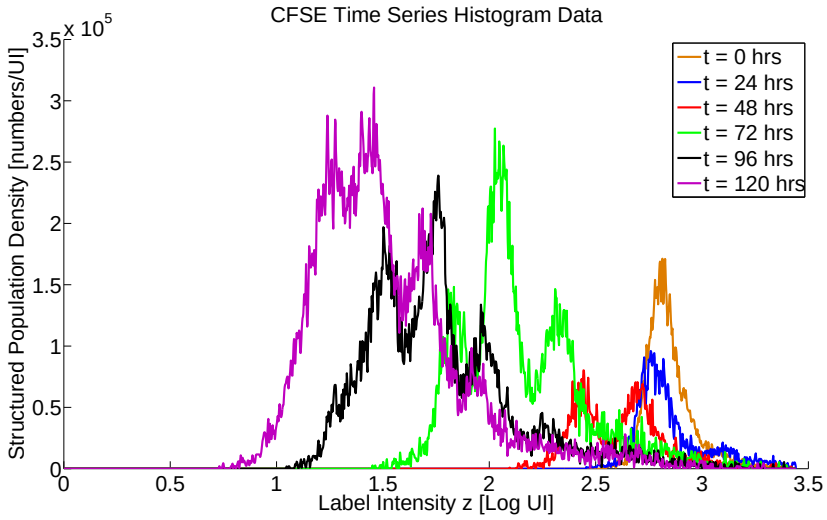

Figure 3.1: Overlap between subsequent peaks in a CFSE histogram data set. While it is easy to distinguish the various peaks in the data, the overlap between peaks results in some systematic error when attempting to identify a region of the horizontal axis with a specific division number. While the slow drift to the left over time (as a result of intracellular turnover of CFSE) can be accounted for using a 'translated coordinate' (Chapter 2), it is unclear how the overlap between subsequent generations of cells affects the estimated proliferation and death rates. Regardless, this overlap also prevents the accurate determination of cell numbers.

can clearly discern the distinct generations of cells in the data set depicted in Figure 3.1. However, the peaks corresponding to particular generations of cells overlap slightly and drift to the left in time (as a result of CFSE turnover), thus weakening the correlation between the state variable and division number. It was shown in Chapter 2 that the proliferation and death rates can be parameterized with respect to a 'translated variable' which accounts for the loss of measured FI in time, and that this translated variable is more strongly correlated with division number than the original structure variable x. Thus it is possible to estimate proliferation and death rate functions $\tilde{\alpha}(t,s)$ and $\beta(\tilde{t},s)$ which can be used to compute average division rates in terms of the number of divisions undergone (see, e.g., Figure 2.14 in Chapter 2).

Still, the overlap between distinct peaks in the data remains problematic. Deconvolution techniques (such as fitting peaks with normal or lognormal curves) impose particular forms on the experimental data which may bias the computed number of cells in each generation. While the model (3.1) is advantageous in being able to estimate average proliferation and death rates without any deconvolution of the data into cell numbers, it cannot be used to accurately assess the number of cells in a particular generation. This information could be approximated by integrating the structured density $n(t,x)$ over a region $[x_1,x_2]$ (corresponding approximately to the location of a given peak in the histogram data), but this approximation is limited by the extent to which distinct generations of cells in the histogram data overlap. It is also not clear how much error may be introduced into the estimated proliferation and death rates by this overlap of distinct generations.

Fortunately, a very simple reformulation of (3.1) permits both the accurate quantification of total cells per division number and the accurate estimation of proliferation and death rates in terms of division number. Rather than modeling the *population* with a single differential equation, one can model each *generation* of cells with a single equation,

$$\frac{\partial n_i}{\partial t} + \frac{\partial [v(t,x) n_i(t,x)]}{\partial x} = -(\alpha_i(t) + \beta_i(t)) n_i(t,x) + R_i(t,x),$$

with the generations linked through the division mechanism $R_i(t,x)$ as a source term (see the next section). This is a common technique in existing ordinary and delay differential equations models for dividing cells (see [38, 39, 34]). Because each generation of cells is assigned to a particular compartment (indexed by i) with unique proliferation and death rates, it is not necessary to estimate these rates in terms of the structure variable x, so that peak overlap and label decay no longer affect the accuracy of the estimated rates. This is in contrast to previous work [20, 21] in which considerable space is devoted to answering the question of how to parameterize the structural dependence of the proliferation and death rates. As an added advantage, the number of parameters necessary for the parameterizations of the proliferation and death rates is reduced (because there is no longer a need to parameterize the functions α_i and β_i in terms of the structure variable). Furthermore, the existence of multiple compartments makes it possible to accurately determine cell numbers in terms of divisions undergone, even though the computed densities (for the distinct compartments) will still overlap when placed simultaneously on the x axis. Because this model does not rely upon any assumptions as to the shape (normal, lognormal, etc.) of the generation peaks (instead starting from an initial condition and fitting directly to the CFSE histogram data) systematic bias should be avoided just as when using the fragmentation equation (3.1).

In this chapter, we begin with a careful derivation of the compartmental model. The solution to this model is then presented and computational aspects are discussed. Next we formulate an inverse problem for the estimation of the AutoFI and Gompertz parameters, as well as the proliferation and death rate functions $\alpha_i(t)$ and $\beta_i(t)$. As in the previous chapter, multiple parameterizations of the proliferation and death rate functions are considered with the goal of determining how these rates depend on both division number and on time. After presenting results which demonstrate the enhanced capabilities of the compartmental model, the statistical properties of the flow cytometry data are considered and ramifications for the quantification of uncertainty in the estimated parameters are discussed.

3.2 Derivation of the Compartmental Model

The derivation of the compartmental model follows immediately from the derivation of the fragmentation model (3.1) presented in Chapter 2, which is itself a variation of the structured population models of Bell-Anderson [26] and Sinko-Streifer [97]. Let $n_i(t,x)$, $0 \le i \le i_{\max}$ be the label structured population density of a population of cells stained with CFSE and having undergone i divisions. The structure variable x is the fluorescence intensity (FI) of a cell (in arbitrary units of intensity, UI) satisfying $x \ge x_a$ where x_a is the natural autofluorescence intensity (AutoFI) of cells; t is time (in hours). While it is known that AutoFI increases significantly when cells become activated, this increase is not believed to be significant for the current modeling effort (as AutoFI contributes minimally to the measured FI

of a labeled but unactivated cell). Thus, the parameter x_a should be understood to describe AutoFI for *activated* cells. It is known that FI scales linearly with the concentration of CFSE used to label a population of cells, and that this measurement does not change significantly when cells increase in size [80]. Thus we assume FI is a mass-like quantity.

For the derivation, we use the structured density to define the total number of cells with fluorescence intensity in an arbitrary region $[x_0, x_1]$ at time t having undergone i divisions:

$$N_i(t) = \int_{x_0}^{x_1} n_i(t,x)dx.$$

As in [20], let $\Delta x(t, x, \Delta t)$ be the average increase of FI of cells with initial intensity x during the interval $(t, t + \Delta t)$ and assume that Δt is chosen such that $|\Delta x| << x_1 - x_0$ (so that the number of cells which move into the region via division and subsequently divide, die, or drift out of the region is negligible). This involves the tacit assumption that all cells, regardless of the number of divisions undergone, lose label in an identical manner which depends only on time and FI. Because cells are observed to only decrease in FI as a result of the natural decay of the fluorescent label (most likely resulting from the metabolism-related turnover of the intracellular proteins to which the dye is bound), it should be noted that Δx will be non-positive. Thus subtraction by Δx actually results in a larger value. While counterintuitive, this definition is maintained in order to harmonize with other structured population models.

We now consider the quantity $N_i(t + \Delta t) - N_i(t)$, the change in $N_i(t)$ during the time interval $(t, t + \Delta t)$. We consider the following five possible contributions:

(i.) Cells with intensity in the interval $[x_1, x_1 - \Delta x(t, x_1, \Delta t)]$, losing FI according to Δx:

$$\int_{x_1}^{x_1 - \Delta x(t,x_1,\Delta t)} n_i(t,x)dx.$$

(ii.) Cells with intensity in the interval $[x_0, x_0 - \Delta x(t, x_0, \Delta t)]$, losing FI according to Δx:

$$\int_{x_0}^{x_0 - \Delta x(t,x_0,\Delta t)} n_i(t,x)dx.$$

(iii.) Cells which would have contributed to $N_i(t + \Delta t)$ had they not died:

$$\int_t^{t+\Delta t} \int_{x_0 - \Delta x(t,x_0,t+\Delta t+\tau)}^{x_1 - \Delta x(t,x_1,t+\Delta t+\tau)} \beta_i(\tau) n_i(\tau,x)dxd\tau.$$

(iv.) The disappearance of cells from the region due to proliferation:

$$\int_t^{t+\Delta t} \int_{x_0 - \Delta x(t,x_0,t+\Delta t+\tau)}^{x_1 - \Delta x(t,x_1,t+\Delta t+\tau)} \alpha_i(\tau) n_i(\tau,x)dxd\tau.$$

(v.) The gain of two daughter cells in the region as a result of proliferation in the parent generation:

$$2\int_t^{t+\Delta t}\int_{2(x_0-\Delta x(t,x_0,t+\Delta t+\tau))-x_a}^{2(x_1+\Delta x(t,x_0,t+\Delta t+\tau))-x_a}\alpha_{i-1}(\tau)n_{i-1}(\tau,x)dxd\tau.$$

This term is justified in greater detail in [20].

It should be noted that component (v.) will not appear for the undivided generation ($i = 0$). The most significant change distinguishing this new model from previous modeling attempts is that the proliferation and death rates (both in units hr^{-1}) for cells having undergone i divisions, α_i and β_i, respectively, are now assumed to be independent of the structure variable x. Previous efforts [20, 21, 75, 77] used the dependence of the functions α and β (which in those documents described proliferation and death rates for the entire population of cells, regardless of the number of divisions undergone) on the structure variable in order to account for the dependence of those rates on division number, which is believed to be necessary [38, 39, 72, 76]. Because the compartmental model explicitly identifies the generation of cells being described by a particular rate function, this dependence is no longer necessary. We continue to permit the dependence of the proliferation and death rates on time, as determining the necessity of doing so is a major goal of this research.

Given these contributions, it follows that

$$\begin{aligned}N_i(t+\Delta t)-N_i(t) = &\int_{x_1}^{x_1-\Delta x(t,x_1,\Delta t)} n_i(t,x)dx - \int_{x_0}^{x_0-\Delta x(t,x_0,\Delta t)} n_i(t,x)dx\\ &-\int_t^{t+\Delta t}\int_{x_0-\Delta x(t,x_0,t+\Delta t+\tau)}^{x_1-\Delta x(t,x_1,t+\Delta t+\tau)}\beta_i(\tau)n_i(\tau,x)dxd\tau\\ &-\int_t^{t+\Delta t}\int_{x_0-\Delta x(t,x_0,t+\Delta t+\tau)}^{x_1-\Delta x(t,x_1,t+\Delta t+\tau)}\alpha_i(\tau)n_i(\tau,x)dxd\tau\\ &+2\int_t^{t+\Delta t}\int_{2(x_0-\Delta x(t,x_0,t+\Delta t+\tau))-x_a}^{2(x_1+\Delta x(t,x_0,t+\Delta t+\tau))-x_a}\alpha_{i-1}(\tau)n_{i-1}(\tau,x)dxd\tau. \qquad (3.2)\end{aligned}$$

Following the standard procedure of dividing by Δt and letting $\Delta t \to 0$, we obtain $\frac{dN_i}{dt}$ on the left side of (3.2). We now treat the right side of the equation term by term. For the first term on the right side, if $n_i(t,x)$ is continuous in t and x (a reasonable assumption), the mean value theorem (MVT) implies that there exists a $\theta \in [x_1, x_1 - \Delta x(t,x_1,\Delta t)]$ such that

$$\int_{x_1}^{x_1-\Delta x(t,x_1,\Delta t)} n_i(t,x)dx = -\Delta x(t,x_1,\Delta t)n_i(t,\theta).$$

Assuming Δx is continuous in Δt (that is, there is no instantaneous label loss) and varies smoothly, we can write

$$\lim_{\Delta t\to 0}\frac{-\Delta x(t,x_1,\Delta t)}{\Delta t}n_i(t,\theta)d\theta = -v(t,x_1)n_i(t,x_1). \qquad (3.3)$$

where we have defined $\frac{dx}{dt} = v(t,x)$, the instantaneous rate of FI change of cells with intensity x and

time t (units UI/hr). Applying the same arguments to the second term,

$$-\int_{x_0}^{x_0-\Delta x(t,x_0,\Delta t)} n_i(t,x)dx = v(t,x_0)n_i(t,x_0). \tag{3.4}$$

In the consideration of the third term of (3.2), define

$$u_{\beta_i}(\tau) = \int_{x_0-\Delta x(t,x_0,t+\Delta t+\tau)}^{x_1-\Delta x(t,x_1,t+\Delta t+\tau)} \beta_i(\tau)n_i(\tau,x)dx.$$

Then if $\Delta x(t,x,\Delta t)$ and $\beta_i(\tau)n(t,x)$ are continuous functions of their variables, so is $u_{\beta_i}(\tau)$ and by the MVT, there exists a $\theta' \in [t, t+\Delta t]$ such that

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} u_{\beta_i}(\tau)d\tau = u_{\beta_i}(\theta').$$

Thus it follows that

$$\lim_{\Delta t\to 0} u_{\beta_i}(\theta') = u_{\beta_i}(t) = \int_{x_0}^{x_1} \beta_i(t)n_i(t,x)dx, \tag{3.5}$$

assuming $\Delta x(t,x,0) = 0$ for all t and x (which follows from the previous assertion regarding the smoothness of Δx in Δt). Using a similar argument for the fourth term of (3.2),

$$\lim_{\Delta t\to 0} u_{\alpha_i}(\theta') = u_{\alpha_i}(t) = \int_{x_0}^{x_1} \alpha_i(t)n_i(t,x)dx, \tag{3.6}$$

where $u_{\alpha_i}(\tau)$ has the obvious definition. Finally, for the last term of (3.2), the same argument along with the change of variables $\xi = (x+x_a)/2$ results in

$$2\lim_{\Delta t\to 0} u_{\tilde{\alpha}_{i-1}}(\theta') = 4\int_{x_0}^{x_1} \alpha_{i-1}(t)n_i(t,2x-x_a)dx. \tag{3.7}$$

Altogether, we can assemble (3.3) - (3.7) to rewrite Equation (3.2)

$$\begin{aligned}\frac{dN_i}{dt} =& -v(t,x_1)n_i(t,x_1) + v(t,x_0)n_i(t,x_0) - \int_{x_0}^{x_1} \beta_i(t)n_i(t,x)dx \\ &- \int_{x_0}^{x_1} \alpha_i(t)n_i(t,x)dx + 4\int_{x_0}^{x_1} \alpha_{i-1}(t)n_{i-1}(t,2x-x_a)dx.\end{aligned} \tag{3.8}$$

On the left side of this equation, by the definition of $N_i(t)$,

$$\frac{dN_i}{dt} = \int_{x_0}^{x_1} \frac{\partial n_i(t,x)}{\partial t}dx.$$

Finally, by applying the Fundamental Theorem of Calculus to the first two terms on the right side of Equation (3.8)

$$-v(t,x_1)n_i(t,x_1) + v(t,x_0)n_i(t,x_0) = -\int_{x_0}^{x_1} \frac{\partial(v(t,x)n_i(t,x)}{\partial x}.$$

Thus, simplifying and rearranging (3.8),

$$\int_{x_0}^{x_1} \frac{\partial n_i(t,x)}{\partial t} + \int_{x_0}^{x_1} \frac{\partial (v(t,x)n_i(t,x)}{\partial x} = \\ -\int_{x_0}^{x_1} (\alpha_i(t) + \beta_i(t))n_i(t,x) + 4\int_{x_0}^{x_1} \alpha_{i-1}(t)n_{i-1}(t, 2x - x_a).$$

Equivalently, because x_0 and x_1 are arbitrary,

$$\frac{\partial n_i(t,x)}{\partial t} + \frac{\partial [v(t,x)n_i(t,x)]}{\partial x} = \\ -(\alpha_i(t) + \beta_i(t))n_i(t,x) + 4\alpha_{i-1}(t)n_{i-1}(t, 2x - x_a). \tag{3.9}$$

Thus it follows that the total population of all cells is modeled by the system of PDEs

$$\begin{aligned} \frac{\partial n_0}{\partial t} + \frac{\partial [v(t,x)n_0(t,x)]}{\partial x} &= -(\alpha_0(t) + \beta_0(t))n_0(t,x) \\ \frac{\partial n_1}{\partial t} + \frac{\partial [v(t,x)n_1(t,x)]}{\partial x} &= -(\alpha_1(t) + \beta_1(t))n_1(t,x) + R_1(t,x) \\ &\vdots \\ \frac{\partial n_{i_{\max}}}{\partial t} + \frac{\partial [v(t,x)n_{i_{\max}}(t,x)]}{\partial x} &= -\beta_{i_{\max}}(t)n_{i_{\max}}(t,x) + R_{i_{\max}}(t,x) \end{aligned} \tag{3.10}$$

where $R_i(t,x) = 4\alpha_{i-1}(t)n_{i-1}(t, 2x - x_a)$ for $1 \le i \le i_{\max}$. Note the assumption that $\alpha_{i_{\max}} = 0$. While there is no mathematical limit to the number of generations which can be computed, experimental data generally exhibits fewer than 10 divisions. In an inverse problem setting (see Section 3.4), the parameter $i_{\max}$ can be easily fixed in advance by simply counting the number of generations which appear in the data. Because it is then known that there are no cells with generation number $i_{\max} + 1$, there must be no proliferation in generation $i_{\max}$, and the model can be simplified by setting $\alpha_{i_{\max}} = 0$. Of course, the process of determining $i_{\max}$ could be automated via model refinement statistical tests, but that seems unnecessary given the ease with which the parameter can be identified from data.

There is an additional mathematical justification for setting $\alpha_{i_{\max}} = 0$. The total quantity of CFSE FI in the population is

$$M(t) = \int_{x_a}^{\infty} x \left(\sum_{i=0}^{i_{\max}} n_i(t,x) \right) dx.$$

Using the definition of $M(t)$ and the system of equations (3.10), we can show that

$$\frac{dM}{dt} = \int_{x_a}^{\infty} v(t,x) \left(\sum_{i=0}^{i_{\max}} n_i(t,x) \right) - \int_{x_a}^{\infty} x \left(\sum_{i=0}^{i_{\max}} \beta_i(t)n_i(t,x) \right) \\ + x_a \int_{x_a}^{\infty} \left(\sum_{i=0}^{i_{\max}} \alpha_i(t)n_i(t,x) \right) - \int_{x_a}^{\infty} x \left(\alpha_{i_{\max}}(t)n_{i_{\max}}(t,x) \right).$$

While the first three terms on the right side of this equation are physically relevant and expected (loss of FI by Gompertz decay, loss of FI by death, and the additive role of AutoFI, respectively) the final

term is not experimentally valid because cells do not recognize a maximum division number after which they must leave the measured population. The requirement that $\alpha_{i_{\max}} = 0$ eliminates this term.

The initial condition must be prescribed for each i,

$$n_i(0, x) = \Phi_i(x). \tag{3.11}$$

It will generally (but not necessarily always) be true that $\Phi_i(x) = 0$ for $i \geq 1$ (that is, all cells in the population are undivided at $t = 0$). These initial condition curves are determined from data taken at $t = 0$ (see Section 3.3.1). The left ($x = x_a$) boundary conditions are the no-flux boundary conditions

$$v(t, x_a)n_i(t, x_a) = 0 \tag{3.12}$$

for all $0 \leq i \leq i_{\max}$. Because the problem is defined on the semi-infinite domain $x \geq x_a$, these conditions are sufficient to compute a solution. (This is in contrast to previous work in Chapter 2 and elsewhere [20, 21, 75, 77] in which a zero-recruitment boundary condition, $n(t, x_{\max}) = 0$, is imposed at the right boundary of the computational domain. Under appropriate conditions, these two formulations are equivalent, as discussed in the next section.)

3.3 Model Solution

For many decay velocities $v(t, x)$ of interest, the system of equations (3.10) can be solved analytically using the method of characteristics. As discussed previously, we assume that the rate at which cells naturally lose FI is described by the same function, $v(t, x)$, for all cells regardless of division number. As such, the characteristic lines are the same for each generation of cells. Furthermore, it will be assumed that this rate of FI loss is adequately described by a Gompertz decay process [69], which has been shown [20] to effectively describe the biphasic decay [84, 89, 104] characteristic of proliferation assay data when the intracellular label is CFSE (see Chapter 2). Thus we have

$$v(t, x) = -c(x - x_a)e^{-kt} \tag{3.13}$$

where $c > 0$ and $k > 0$ (both with units 1/hr) are parameters to be determined. In effect, this function describes cellular FI which decreases exponentially (with initial rate c) to the level of cellular AutoFI, while the exponential rate itself decreases (exponentially) with rate k. The assumption of Gompertz decay of cellular FI has the additional benefit of trivially satisfying the left boundary condition (3.12) for all i, provided $n_i(t, x_a)$ is finite (so that the flux at the boundary is well-defined).

Incorporating the Gompertz decay process, the system (3.10) can be rewritten

$$\begin{aligned}
\frac{\partial n_0}{\partial t} - ce^{-kt}(x-x_a)\frac{\partial n_0}{\partial x} &= -(\alpha_0(t)+\beta_0(t)-ce^{-kt})n_0(t,x) \\
\frac{\partial n_1}{\partial t} - ce^{-kt}(x-x_a)\frac{\partial n_1}{\partial x} &= -(\alpha_1(t)+\beta_1(t)-ce^{-kt})n_1(t,x) + R_1(t,x) \\
&\vdots \\
\frac{\partial n_{i_{\max}}}{\partial t} - ce^{-kt}(x-x_a)\frac{\partial n_{i_{\max}}}{\partial x} &= -(\beta_{i_{\max}}(t)-ce^{-kt})n_{i_{\max}}(t,x) + R_{i_{\max}}(t,x).
\end{aligned} \tag{3.14}$$

The characteristic lines (for all i) are described by

$$\frac{dx}{dt} = v(t,x) = -c(x-x_a)e^{-kt}, \tag{3.15}$$

and hence the characteristic line emanating from the point $(0,s)$ in the tx-plane is

$$x(t;s) = x_a + (s-x_a)\exp\left[-\frac{c}{k}\left(1-e^{-kt}\right)\right], \tag{3.16}$$

where $s \geq x_a$ parameterizes the line along which the initial condition is prescribed.

Define

$$f_i(t) = \alpha_i(t) + \beta_i(t) - ce^{-kt}.$$

For undivided cells ($i=0$), the solution along a characteristic line emanating from a point $(0,s)$ in the tx-plane is given by

$$\frac{\partial n_0}{\partial t} = -f_0(t)n_0(t,x(t;s))$$

with $n_0(0,x(0;s)) = n_0(0,s) = \Phi_0(s)$. Thus the solution along characteristic lines is

$$n_0(t,x(t;s)) = \Phi_0(s)\exp\left(-\int_0^t f_0(\tau)d\tau\right). \tag{3.17}$$

As written above, the system of equations (3.14) is defined on the semi-infinite domain $x \geq x_a$. In general, the initial condition function $\Phi_0(x)$, can be determined from data (see Section 3.4) only on some finite segment $[x_a, x_{\max}]$ of the domain. However, there is no loss of generality in extending the initial condition curve by assuming $\Phi_0(x) = 0$ if $x > x_{\max}$. This is in contrast to the previous chapter (and other label structured models [20, 21, 75, 77]) in which a PDE was defined only on the finite interval $[x_a, x_{\max}]$ and a zero-recruitment boundary was imposed. In fact, the two formulations are equivalent provided $\Phi(x_{\max}) = 0$ (in the former models; $\Phi_i(x_{\max}) = 0$ for all i in the compartmental model) and $v(t,x) < 0$. As the semi-infinite formulation is notationally simpler and easy to implement, we use it here.

The solutions for $i \geq 1$ along the same characteristic lines (3.16) are described by

$$\frac{\partial n_i}{\partial t} = -f_i(t)n_i(t,x(t;s)) + R_i(t,x(t;s)) \tag{3.18}$$

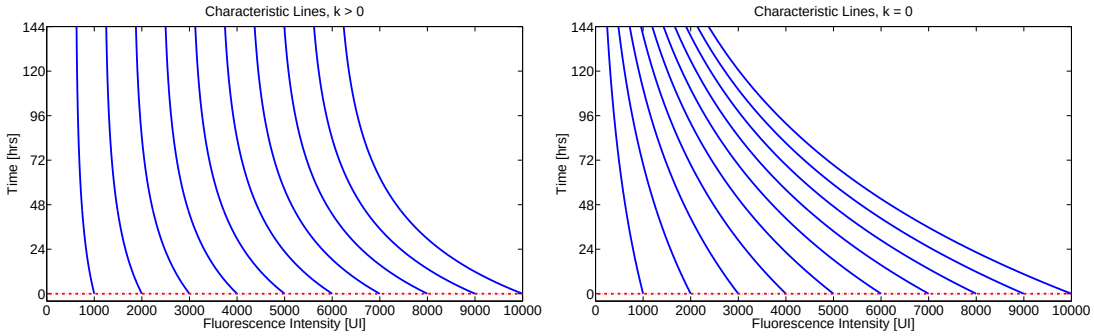

Figure 3.2: Characteristic lines given by Equations (3.15)-(3.16) when $c = 1 \times 10^{-2}$ and $k = 2 \times 10^{-2}$ (left) and in the limiting case when $k = 0$ (right). Notice that distinct characteristic lines will remain separated by some positive distance for all time in the former case, while in the latter case the lines asymptotically converge to $x = x_a$. The broken line along the bottom of both graphics is the line along which the initial condition data is given. It is clear that this initial condition curve is nowhere tangent to a characteristic line, hence the local existence of a unique solution.

with $n_i(0, x(0; s)) = \Phi_i(s)$ and the solutions are

$$n_i(t, x(t; s)) = \Phi_i(s) \exp\left(-\int_0^t f_i(\tau) d\tau\right) + \int_0^t R_i(\tau, x(\tau; s)) \exp\left(-\int_\tau^t f_i(\xi) d\xi\right) d\tau. \tag{3.19}$$

It is worth noting that the solution by the method of characteristics involves the construction of an integral surface in the coordinates t and s. The change of coordinates from t and x to t and s has Jacobian

$$J = \begin{vmatrix} \frac{\partial t}{\partial t} & \frac{\partial t}{\partial s} \\ \frac{\partial x}{\partial t} & \frac{\partial x}{\partial s} \end{vmatrix} = \exp\left(-\frac{c}{k}\left(1 - e^{-kt}\right)\right),$$

which is nonsingular along the initial condition curve ($t = 0$). Hence we are guaranteed (by the construction above) that a unique solution exists at least locally near the initial condition curve. Note that in the limit as $k \to 0^+$, the Jacobian is $J_{k \downarrow 0} = e^{-ct}$, which becomes singular asymptotically in time (reflecting the asymptotic convergence of the characteristic lines). In such a case, one might observe solutions which grow without bound. This is only of minimal concern, however, as the total label loss resulting from decay is small over the duration of a typical experiment. Possible characteristics lines (for $k > 0$ and $k = 0$) are shown graphically in Figure 3.2.

For the remainder of this document, it will be assumed that all cells are undivided at $t = 0$, so that $\Phi_i(x) = 0$ for $i \geq 1$. This condition is satisfied by essentially all experimental data. Thus, the only nontrivial initial condition for the PDE system (3.14) is $\Phi_0(x)$. As this model is motivated by an attempt to fit and explain experimental data, this smooth initial condition must be constructed from data taken at the beginning of the experiment. Our process for doing so is described below, followed by the numerical algorithm for computing the solutions (3.17) and (3.19).

3.3.1 Initial Condition Construction

Just as in the previous chapter, we use experimental data (which is noisy histogram data in the logarithmic coordinate $z = \log_{10}(x)$) collected at $t = 0$ hours in order to determine $\Phi_0(x)$. The data consist of ordered pairs (z_k^0, n_k^0), which denote the number of cells n_k^0 counted into the histogram bin (subject to measurement error) with its left boundary at z_k^0 when $t = 0$ (see Section 1.2 for more details). In order to obtain a smooth initial condition function from the noisy data (z_k^0, n_k^0), a smooth line is drawn through the original histogram data which is taken to represent the 'true' cell counts in the absence of noise. The numerical values are recovered from the smooth line using DataThief [101] to form the 'noiseless' counts $(z_k^0, \hat{n}_k^0)$, which are then easily transferred from the logarithmic coordinate resulting in new ordered pairs $(x_k^0, \hat{n}_k^0)$ (because the $\hat{n}_k^0$ are approximate numbers of counted cells as opposed to a structured density, the values do not need to be rescaled when changing from z to x).

Finally, we must use these 'noiseless' cell counts in the x coordinate in order to determine the structured density initial condition $\Phi_0(x)$ for (3.17). To do so, we first define the function

$$\varphi(x) = \sum_k \hat{n}_k^0 l_k(x),$$

where $l_k(x)$ are piecewise linear functions satisfying

$$l_j(x_k^0) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}.$$

Thus the function $\varphi(x)$ is a piecewise linear function such that $\varphi(x_k^0) = \hat{n}_k^0$. Next, we compute the total measured FI in the population at $t = 0$ using the original noisy data,

$$FI_{data} = \sum_k x_k^0 n_k^0.$$

Similarly, the total FI in the smooth data function $\varphi(x)$ is

$$FI_{smooth} = \int x\varphi(x)dx,$$

where the integral is approximated using the composite trapezoidal rule. The initial condition function is then constructed as

$$\Phi_0(x) = \frac{FI_{data}}{FI_{smooth}}\varphi(x).$$

The results of this technique are shown in Figure 3.3. This method ensures that the total measured FI in the initial condition curve is equal to the total measured FI in the original noisy data. Because the mathematical model (3.9) is derived from conservation principles (considering FI as a mass-like quantity), this provides a useful comparison between the data and the model, as well as a method to assess the accuracy of the numerical simulations. It is worth noting that such a complex procedure is unnecessary in the event the histogram bins are evenly spaced (in the logarithmic coordinate z). In such an event, a smooth density function (in z) can be computed from the smooth cell counts $(z_k^0, \hat{n}_k^0)$ simply

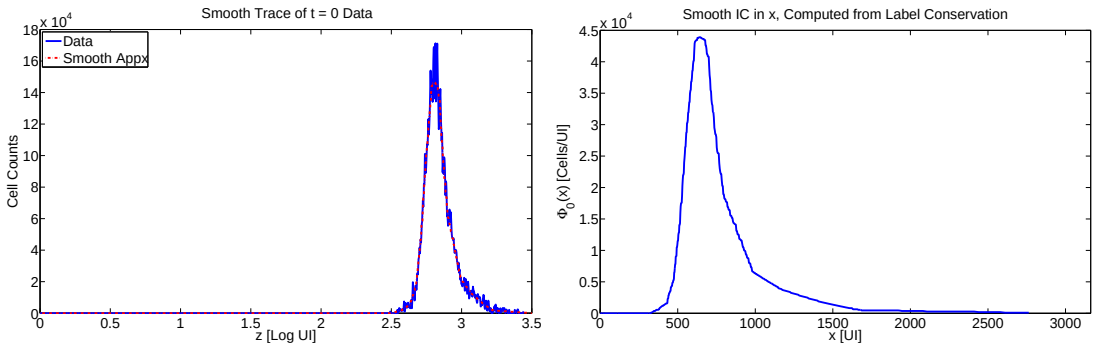

Figure 3.3: Left: Smooth curve drawn through the experimental data taken at $t = 0$ hours. Right: Initial condition function $\Phi_0(x)$ computed from the smooth line using the algorithm of Section 3.3.1.

by dividing the counts by the bin spacing. The function $\Phi_0(x)$ can then be computed as a simple change of variables from z to x. However, this method of computation may result in discontinuous jumps in the computed density if there are abrupt changes in the sizes of adjacent histogram bins. Moreover, total measured FI for the initial condition curve will not necessarily be equal to the total FI in the noisy data (although the two values should still be close) if such a method were to be used. Thus, we find the rescaling method above to be preferable.

3.3.2 Numerical Solution

Given the initial condition function $\Phi_0(x)$ as computed above, it now remains to numerically compute the solutions (3.17) and (3.19) for $n_0(t, x)$ and $n_i(t, x)$, $1 \leq i \leq i_{\max}$, respectively. In the structure variable, the solutions are computed on a fixed (i.e., one that does not change with division number) mesh $\{x^{(k)}\}$, $1 \leq k \leq N_x$ (these should be distinguished from the $x_k^j = 10^{z_k^j}$ used to describe the data). While it is not strictly necessary for each compartment to be computed on the same grid, there seems to be little advantage in varying the structure variable mesh with division number. Because the major features (the 'peaks') of the structured density solution shrink by approximately a factor of two with each division, it is advantageous to choose the points $\{x^{(k)}\}$ so that they are logarithmically spaced (that is, so that the collection $\{\log_{10}(x^{(k)})\}$ is evenly spaced). This ensures an increasing density of nodes as x decreases, and hence as the major features of the solution become more condensed. The uneven spacing of the nodes for the structural variable does not cause any numerical difficulties as the algorithm presented below requires only interpolation (i.e., no finite-difference derivatives) in the structural dimension. Because the model presented in this report uses distinct compartments for each generation of cells, and because each generation of cells remains in a relatively small region in the structure variable (see [20, 21, 75, 77], which were motivated by this fact), it is certainly true that the use of a single structural mesh for all compartments results in some unnecessary storage and computations. However, the value of N_x has only a small effect on computational time (see below) so that this is of little concern.

In time, the solutions are computed on a fixed, evenly spaced mesh $\{t^{(m)}\}$, with spacing h_t. Unlike time-stepping finite difference methods (such as the Lax-Wendroff method used in [20, 21, 77] and

Chapter 2) which require storage of the solution at only the most recent time steps, the method of characteristics solution (3.19) requires an integration along characteristic lines over the history of the solution (see Equation (3.21) below). This integration is computed via the trapezoidal rule, using quadrature nodes which correspond to the time mesh $\{t^{(m)}\}$. While this method of computing the solution is storage-intensive, the requirements are not unreasonable, even when running in MATLAB on a 32-bit desktop machine.

It is obvious from Equations (3.17) and (3.19) that the system (3.9) can be solved inductively on i. For time $t^{(m)}$ and FI $x^{(k)}$, we find from Equation (3.16)

$$s(t^{(m)}, x^{(k)}) = (x^{(k)} - x_a)\exp\left(\frac{c}{k}\left(1 - e^{-kt^{(m)}}\right)\right) + x_a. \tag{3.20}$$

The solution $n_0(t, x)$ (Equation (3.17)) is then computed by multiplying the values of $\Phi_0(s)$ by the scalar

$$\exp\left(-\int_0^{t^{(m)}} f_0(\tau)d\tau\right) = \exp\left(-\int_0^{t^{(m)}} \left(\alpha_0(\tau) + \beta_0(\tau) - ce^{-k\tau}\right) d\tau\right).$$

For all parameterizations of the functions $\alpha_i(t)$ and $\beta_i(t)$ considered in this report, the integral above can be computed exactly.

As noted above, it is assumed $\Phi_i(x) = 0$ for $i \geq 1$. Thus the solutions $n_i(t, x)$ can be rewritten

$$n_i(t, x) = \int_0^t G_i(\tau; t, x)d\tau \tag{3.21}$$

where

$$G_i(\tau; t, x) = 4\alpha_{i-1}(\tau)n_{i-1}(\tau, 2x(\tau, s) - x_a)\exp\left(-\int_\tau^t f_i(\xi)d\xi\right).$$

As above, the scalar $4\alpha_{i-1}(\tau)\exp\left(-\int_\tau^t f_i(\xi)d\xi\right)$ is computed exactly. The values of the function $n_{i-1}(t, x)$, though already computed, will only be available at discrete points $(t^{(m)}, x^{(k)})$ and thus (3.21) must be computed via quadrature. Because every solution $n_i(t, x)$ is computed on the same, evenly spaced time mesh, a simple solution is to use this same time mesh (with the trapezoidal rule) in order to approximate the integral. Thus, given a point $(t^{(l)}, x^{(k)})$, $l \leq m$, we must first determine s according to (3.20). This is then used to compute

$$\tilde{x}^{(l)} = 2x(t^{(l)}, s) - x_a,$$

for $0 \leq l \leq m$, where $x(t, s)$ is given in Equation (3.16). Thus we have

$$G_i(t^{(l)}; t^{(m)}, x^{(k)}) = 4\alpha_{i-1}(t^{(l)})n_{i-1}(t^{(l)}, \tilde{x}^{(l)})\exp\left(-\int_{t^{(l)}}^{t^{(m)}} f_i(\xi)d\xi\right),$$

with the values of $n_{i-1}(t^{(l)}, \tilde{x}^{(l)})$ determined by linear interpolation. Finally, from (3.21),

$$\begin{aligned} n_i(t,x) = &\int_0^t G_i(\tau;t)d\tau \approx \\ &= h_t \cdot \sum_{l=1}^{m-1} G_i(t^{(l)};t^{(m)},x^{(k)}) + \frac{h_t}{2}\left(G_i(t^{(0)};t^{(m)},x^{(k)}) + G_i(t^{(m)};t^{(m)},x^{(k)})\right). \end{aligned} \tag{3.22}$$

We now consider how the computed solution changes as the mesh parameters N_x and h_t are changed. As a test case, nominal parameters (for x_a, c, k, $\alpha_i(t)$ and $\beta_i(t)$) were used to compute a solution at $t = 120$ hours using various combinations of values for N_x and h_t. The results are shown in comparison in Figure 3.4. For convenience, the solutions $n_i(t,x)$ have been summed together and graphed in terms of the log FI ($z = \log_{10}(x)$) coordinate.

As noted above, the algorithm does not require any quadrature or finite differences in the structural component. At most, it is necessary to use interpolation (linear interpolation seems sufficient) in order to compute the function G_i above in the likely event $\tilde{x}^{(l)} \notin \{x^{(k)}\}$. Thus, one would expect approximately second-order accuracy in N_x. In fact, we find (computationally) this expectation is exceeded. The explanation lies in the iterative manner in which the solution is computed. Consider computing $n_1(t,x)$ (Equation (3.21)), provided $n_0(t,x)$ is already computed. On one hand, because this computation will require (linear) interpolation, we would expect the resulting error to depend upon the mesh-spacing of $n_0(t,x)$. However, $n_0(t,x)$ is computed from the initial condition function $\Phi_0(x)$ which is defined (see Section 3.3.1) as a piecewise linear function. It follows that, as N_x approaches the number of points used in defining $\Phi_0(x)$, the error will no longer decrease (because a piecewise linear function is being used to approximate a more coarsely defined piecewise linear function). In this report, the function $\Phi_0(x)$ is defined with 806 points. Thus, it is no surprise that we find little difference in the solutions computed with $N_x = 512$ and $N_x = 1024$.

The use of the trapezoidal rule with step size h_t to approximate the integral in (3.21) results in a numerical solution which is second order in h_t. Also, we see that, at each time step $t^{(m)}$ ($0 \le m \le T/h_t$, for a solution computed on $t \in [0,T]$), Equation (3.22) requires m computations of the function $G_i(\tau;t,x)$. Thus we expect the computational time to scale as $O(1/h_t^2)$. Table 3.1 summarizes the average computational time for various combinations of N_x and h_t. As expected, computational time approximately quadruples as h_t is halved. As the algorithm in the previous section has been fully vectorized, N_x has only a minimal effect on computational time.

When fitting the compartmental model to data in an inverse problem setting, we must balance the need for an accurate solution with the desire to quickly evaluate the model (given a set of parameters). As such, in the results presented in Section 3.5, we use $N_x = 512$ with $h_t = 0.5$ hours.

3.4 Inverse Problem Formulation

We now consider the inverse problem of calibrating the model (3.14) to a particular data set. As stated in Chapter 1, the data consist of ordered pairs (z_k^j, n_k^j), indicating the total (i.e., after scaling to account for the fraction of cells actually measured) number of cells n_k^j counted into the histogram bins with left

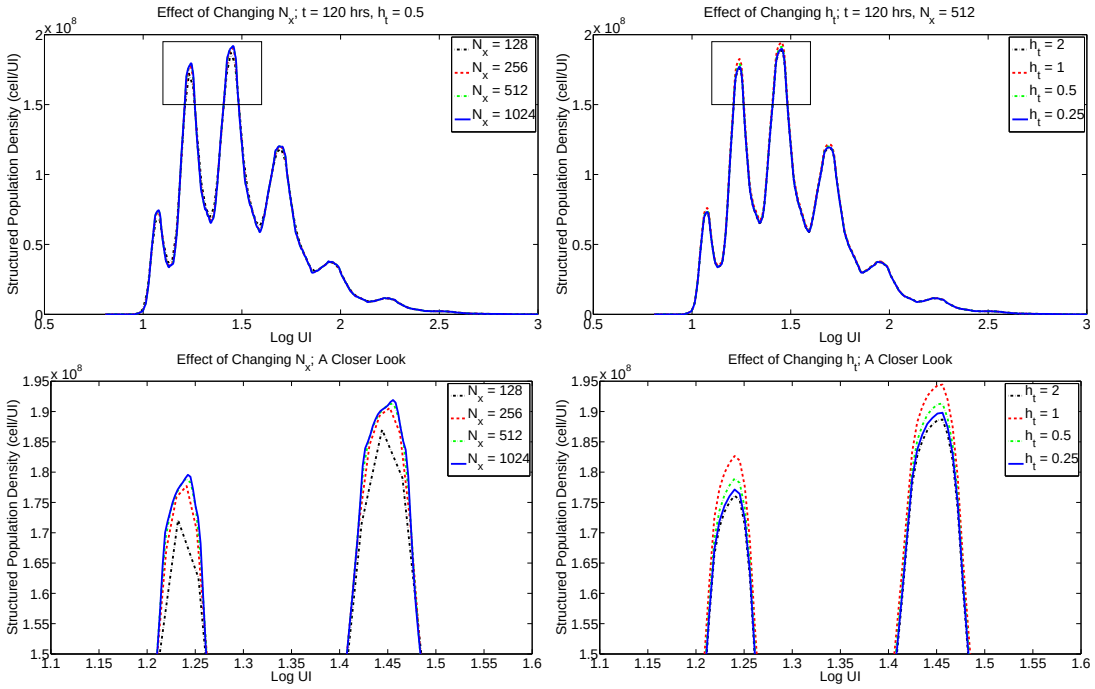

Figure 3.4: Left: Effect of changing the number of structure variable nodes N_x with the time increment fixed at $h_t = 0.5$ hours. The computed solutions are similar for all values of N_x shown (top). Zooming in (bottom), we find that there is only a small difference (less than 2% max) in the computed solutions for $N_x \geq 256$ with solutions for $N_x = 512$ and $N_x = 1024$ virtually indistinguishable. Right: Effect of changing the time increment h_t with the number of structural variable nodes fixed at $N_x = 512$. While the difference between the solution computed for $h_t = 2$ and $h_t = 1$ is large, there is a much smaller difference (approximately 1% max) between $h_t = 0.5$ and $h_t = 0.25$, as expected. Note that the proximity of the $h_t = 2$ solutions to the $h_t = 0.25$ solution is mere coincidence and does not hold more generally.

Table 3.1: Effects of h_t and N_x on computational time. Computational times are shown in seconds, with h_t specified in hours. As expected, computational time is quadratic in h_t. Meanwhile, N_x has a much smaller effect on computational time.

$N_x \setminus h_t$	2.00	1.00	0.50	0.25
128	2.1	7.5	29.8	120.0
256	2.3	8.9	35.6	150.7
512	2.7	10.9	44.8	199.7
1024	3.9	16.8	69.7	297.9

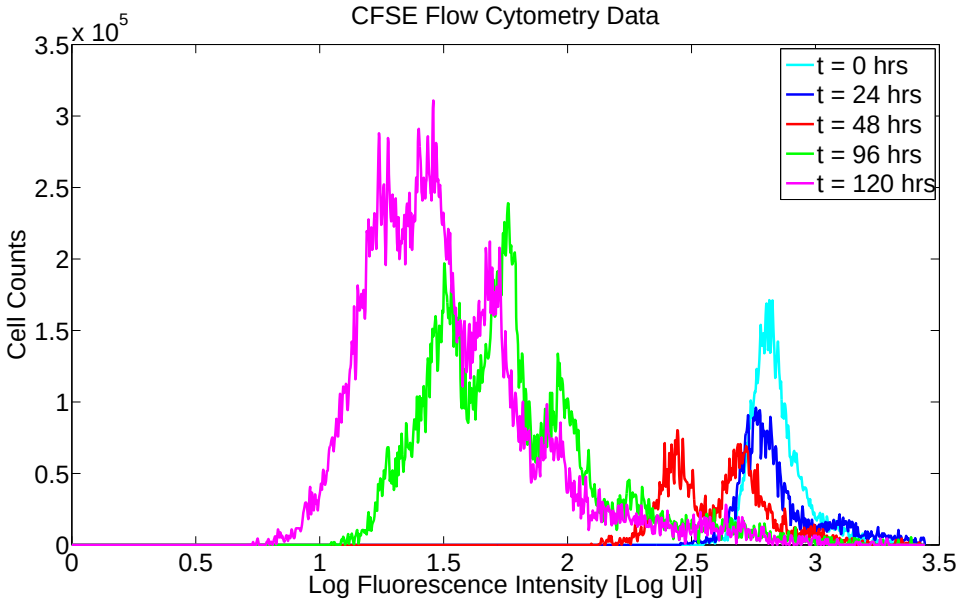

Figure 3.5: CFSE data set for the compartmental model.

boundary at z_k^j (in the log FI coordinate) at time t_j. The notation is meant to emphasize the possibility that the histogram bins need not share a common fixed width, nor need they be the same at each measurement time. The data set we will use to calibrate the compartmental model is shown in Figure 3.5, with measurements taken at $t = 24$, 48, 96, and 120 hours. We proceed in a similar manner to Chapter 2, establishing a mathematical basis for comparing the numerical solution of the compartmental model to this data.

Let $n_i(t, x)$ be the solution of the compartmental model for cells having undergone i divisions. Then the total population of cells is

$$n(t, x) = \sum_{i=0}^{i_{\max}} n_i(t, x).$$

Because this model solution is computed in the linear FI coordinate x while the data is given in the logarithmic FI coordinate $z = \log_{10}(x)$, we define

$$\tilde{n}(t, z) = 10^z \ln(10) n(t, x(z)) = 10^z \ln(10) n(t, 10^z). \tag{3.23}$$

The function $\tilde{n}(t, z)$ is the structured population density in terms of the new structure variable z. The factor $10^z \ln(10)$ arises from the chain rule in the integral formulation of the model (see Section 3.2) and is needed to conserve the quantity of label after the change of variables. Finally, we need to convert this structured density into cell counts for comparison with the data. Thus we define

$$I[\tilde{n}](t_j, z_k^j) \equiv \int_{z_k^j}^{z_k^{j+1}} \tilde{n}(t_j, z) dz,$$

which is the observation operator for the compartmental model. In practice, because the transformed model solution $\tilde{n}(t,z)$ is computed only at discrete points (t_j, z_k^j), we must approximate this observation operator,

$$I[\tilde{n}](t_j, z_k^j) \approx I_A[\tilde{n}](t_j, z_k^j) = \left[\frac{\tilde{n}(t_j, z_k^{j+1}) + \hat{n}(t_j, z_k^{j+1})}{2}\right]\left(z_k^{j+1} - z_k^j\right). \tag{3.24}$$

3.4.1 Ordinary Least Squares

Given an initial condition as constructed in Section 3.3.1, the solution $n(t,x)$ (and hence, $\tilde{n}(t,z)$) is completely determined by the parameters x_a (AutoFI), c and k (Gompertz decay), as well as the proliferation rates $\{\alpha_i(t)\}$ and the death rates $\{\beta_i(t)\}$. Let $\theta = \{x_a, c, k, \{\alpha_i(t)\}, \{\beta_i(t)\}\} \subset \Theta$, where Θ is some set of admissible values for θ. (While it will be necessary to make some simplifying assumptions on Θ in order to make the inverse problem computationally tractable, we postpone that discussion for the moment and proceed with a general overview of the inverse problem procedure.) Thus we may write the model as $n(t,x;\theta)$. The goal of the inverse problem is to determine some value of the parameter θ which minimizes the distance (in an appropriate sense) between the cell counts determined by the model solution, $I[\tilde{n}](t_j, z_k^j)$, and the histogram data. For this report, we choose least squares as the method of estimation. Following standard inverse problem procedure [22, 35, 37, 96], we define the random variables

$$N_k^j = I[\tilde{n}](t_j, z_k^j; \theta_0) + \mathcal{E}_{kj}, \tag{3.25}$$

where $\mathcal{E}_{kj}$ are independent random variables satisfying $E[\mathcal{E}_{kj}] = 0$ representing measurement error and/or 'noise' in the data. The parameter θ_0 is the 'true' parameter (given the model) which is assumed to exist and to describe the data. The data, then, represent a single realization of these random variables,

$$n_k^j = I[\tilde{n}](t_j, z_k^j; \theta_0) + \epsilon_{kj}.$$

The assumption that the data are generated from the specified model, given a nominal truth parameter, is common in inverse problem formulations [8, 22]. While θ_0 is generally unknown, we can define the estimator

$$\theta_{WLS} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{R_{kj}^2}{w_{kj}} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{1}{w_{kj}} \left(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j\right)^2, \tag{3.26}$$

which minimizes the weighted sum (with weights w_{kj}^{-1}) of squared residuals R_{kj}. Because the N_k^j are random variables, so are the R_{kj} and, hence, so is θ_{WLS}. Using the data, we may obtain the estimate

$$\hat{\theta}_{WLS}(n_k^j) = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{r_{kj}^2}{w_{kj}} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{1}{w_{kj}} \left(I[\tilde{n}](t_j, z_k^j; \theta) - n_k^j\right)^2.$$

In theory, the weights w_{kj}^{-1} should be chosen to reflect the variance of the random variables N_k^j. In fact, the accurate, unbiased estimation of standard errors as well as confidence intervals around parameter estimates is premised upon an accurate statistical model (hence accurate weights) for the error terms $\mathcal{E}_{kj}$. In practice, however, such a statistical model is rarely (if ever) known a priori, and

some additional assumptions must be made. For this report, we assume a constant variance (CV) error model, $Var(\mathcal{E}_{kj}) = \sigma_0^2$ for all k and j. In this case, $w_{kj} = 1$ for all k and j and (3.26) becomes an ordinary least squares (OLS) problem,

$$\theta_{OLS} = \arg\min_{\theta\in\Theta} J(\theta|N_k^j) = \sum_{k,j} \mathcal{R}_{kj}^2 = \arg\min_{\theta\in\Theta} \sum_{k,j} \left(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j\right)^2, \tag{3.27}$$

with corresponding estimate

$$\hat{\theta}_{OLS}(n_k^j) = \arg\min_{\theta\in\Theta} J(\theta|n_k^j).$$

The function $J(\theta|n_k^j)$ is the OLS cost of the model, given the data, and is often written simply as $J(\theta)$. The expanded notation is meant to emphasize the dependence of the estimate on the particular data set used to fit the model.

It should be noted that, rather than consider constant variance errors in an OLS framework, one could alternatively consider a statistical model with constant coefficient of variation (CCV), $Var(\mathcal{E}_{kj}) = \sigma_0^2(I[\tilde{n}](t_j, z_k^j; \theta_0))^2$. Then $w_{kj} = (I[\tilde{n}](t_j, z_k^j; \theta_{GLS}))^2$ and (3.26) becomes the generalized least squares (GLS) problem defined implicitly by

$$\theta_{GLS} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{\mathcal{R}_{kj}^2}{(I[\tilde{n}](t_j, z_k^j; \theta_{GLS}))^2} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{\left(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j\right)^2}{(I[\tilde{n}](t_j, z_k^j; \theta_{GLS}))^2}, \tag{3.28}$$

with corresponding estimate $\hat{\theta}_{GLS}(n_k^j)$. As noted above, the results presented in Section 3.5 will focus on parameter estimation in an OLS framework. A more thorough discussion of the reliability of the assumptions for the statistical error model in the inverse problem is postponed until Chapter 4. For the moment, we focus on the applicability of the compartmental model to a particular data set–that is, how well the compartmental model fits the data. Of course, the measure of fit is assessed in an OLS framework, which may be slightly different than a GLS or more general WLS framework. The misspecification of the error model is known to result in biased standard errors (and hence confidence intervals), and thus no such work is carried out here. In spite of this drawback, a slight misspecification of the exact error model should have only minimal effect on the estimated best-fit parameters (see, e.g., the computational example of Section 3.4.2 of [22]), and thus we proceed with the OLS estimation of θ_0.

3.4.2 Parameterizations of Proliferation and Death Rates

We have already defined the parameter $\theta = \{x_a, c, k, \{\alpha_i(t)\}, \{\beta_i(t)\}\} \subset \Theta$ which describes a given model solution. The parameters x_a, c, and k are all elements of $\mathbb{R}$ (although see below, where we consider a probability distribution over the parameter x_a) and thus pose no problem for the estimation procedure. However, the proliferation and death rates $\alpha_i(t)$ and $\beta_i(t)$, $0 \le i \le i_{\max}$ are contained in some (infinite-dimensional) function space. Mathematically, the solutions (3.17) and (3.19) require only $\alpha_i(t)$, $\beta_i(t) \in L_2(0,T)$ in order for the solution to be well-defined. Because it can reasonably be assumed that these functions are bounded, this condition is naturally met. However, as currently written, (3.27) contains a minimization over an infinite-dimensional space Θ. In order to make the estimation problem suitable to computation, additional assumptions and/or approximations are necessary.

The primary motivation for using a label-structured PDE model to analyze histogram data from CFSE-based proliferation assays was an attempt to use measured FI as a surrogate for division number and hence to investigate how the proliferation and death rates for a population of cells change with division number. In earlier efforts [20, 21, 75, 77], this was accomplished by allowing the proliferation and death rates to depend explicitly on the state variable (x or z). For the compartmental model formulated in this report, the number of divisions undergone is accounted for directly, so that it is no longer necessary to have the α_i and β_i dependent upon the structure variable (following the assumption that the interference of CFSE with the intracellular machinery is negligible). Additionally, it was found in [20, 21] that explicit time-dependence of the rate of cell proliferation is a significant feature of an accurate label structured PDE model. We would like to continue this investigation using the new compartmental model. Thus, as we consider possible parameterizations of the proliferation and death rate functions, we do so with an eye toward determining the heterogeneity of the rates (that is, how they vary with division number), as well as the possible time-dependence of the proliferation rates.

We begin with the death rate functions $\beta_i(t)$. It has long been observed that a significant proportion of undivided cells die in the first few days in culture, and that this cell death occurs independent of cellular activation [51]. Beyond these observations, we would like to explore how the death rate of cells changes with division number. Thus we consider the following possible parameterizations for the death rate functions $\beta_i(t)$:

B1 $\beta_i(t) = 0$ for all i and for all t;

B2 $\beta_i(t) = \beta$ for all i and for all t;

B3 $\beta_0(t) = \beta_0$, $\beta_i(t) = 0$ for $i \geq 1$;

B4 $\beta_0(t) = \beta_0$, $\beta_i(t) = \beta$ for $i \geq 1$;

B5 $\beta_i(t) = \beta_i$ for each i.

The possibility **B1** is included as a baseline for comparison, as a means of concluding the necessity of a death term in the mathematical model. As noted above, it is expected that a model which lacks a mechanism to describe cell death will predict far to many cells in the population (compared to the experimental observations) [40, 51]. Parameterization **B2** assumes a constant death rate in the population for all cells regardless of division number. Gett and Hodgkin have shown that parameterization **B3**, in which undivided cells die but all cells which proceed through the first division will remain in the population indefinitely, can be accurately used to predict the number of cells in the population up to approximately 90 hours. More generally, one might consider that cells which have divided at least once may die, but at a rate which is possibly different from the rate for undivided cells. This parameterization (**B4**) has also been successfully used to model proliferation assay data [38, 39, 40]. Finally, we consider the possibility that the death rate is completely heterogenous with respect to division number (parameterization **B5**, [38, 49, 60, 71]).

While the model is derived in sufficiently general terms to include time-dependent death rate functions, we do not consider any such parameterizations in this report. It is certainly possible that, for

particular cell lines and under particular culture conditions, feedback mechanisms such as activation-induced cell death may in fact be time-dependent [51]. For a (hypothetical) population of cells which divides almost synchronously, such time-dependence would be identical to division-number-dependence (i.e., a mechanism which does not appear until, say, 90 hours could be equivalently modeled as a mechanism which does not appear until 3 divisions have been completed). Thus it seems reasonable to conclude that, to some extent, the necessity of time-dependent death rates in the mathematical model will depend on the degree of synchronicity observed in the experimental data. At the very least, past experience [20, 21] as well as the results presented here (Section 3.5) seem to indicate little need for such time-dependence, at least for the current data set.

Unlike for the death rate functions, past experience [20, 21] does indicate a potential need for explicit time-dependence of the proliferation rate functions $\alpha_i(t)$. (The fact that time dependence for cell death rates seems to be sufficiently modeled with only division dependence while a similar result does not hold for the proliferation rates may be explained if the time-dependence of the proliferation rates occurs on a scale faster than the average time a cell takes between subsequent divisions.) Thus we consider the following possible parameterizations for the proliferation rate functions:

A1 $\alpha_0(t) = \alpha_0$; $\alpha_i(t) = \alpha$ for all i;

A2 $\alpha_i(t) = \alpha_i$ for all t;

A3 $\alpha_0(t) = \alpha_0 \chi_{[t>t^*]}$; $\alpha_i(t) = \alpha$ for all i;

A4 $\alpha_0(t) = \alpha_0 \chi_{[t>t^*]}$; $\alpha_i(t) = \alpha_i$;

A5 piecewise linear functions of time (see below).

Previous authors [40, 51, 57] have emphasized a special importance for the time required for a cell to complete its first division. In case **A1**, it is assumed that undivided cells divide at a rate which may be different than the rate for divided cells, but that neither of these two rates depends on time [39]. Alternatively, we consider the more general case **A2** where each generation of cells divides with its own (time-independent) rate [72]. We also consider a simple time-dependent mechanism in which there is a delay before cells begin to divide. A quick glance at the data (Figure 3.1) reveals that no division occurs in the population for at least the first 24 hours. Such a delay can be easily incorporated into the model with a step function at some specified time t^*. Previous models [39] have found such a transient in the undivided population to be a significant feature of an accurate mathematical model. The proliferation rates for subsequent generations may (**A4**) or may not (**A3**) vary with the number of divisions undergone.

Finally, following the example of [20], we consider using piecewise linear splines to incorporate time-dependence into the proliferation rates. Given a fixed set of nodes $\{t_{\alpha_i}^{(q)}\}$, we have

$$\alpha_i(t) = \sum_q a_i^{(q)} l_i^{(q)}(t),$$

where $l_i^{(q)}(t_{\alpha_i}^{(p)}) = 1$ if $p = q$ and is zero if $p \neq q$. Table 3.2 shows the nodes $\{t_{\alpha_i}^{(q)}\}$ used for the estimation

Table 3.2: Chosen nodes for the estimation of piecewise linear proliferation rates. Bold font indicates a node for which the proliferation rate was set to zero rather than estimated. For each generation, the proliferation rate is assumed to be zero outside the set of nodes shown. Thus the proliferation rate is estimated at three nodes for each division number.

Generation (i)	$\{t_{\alpha_i}^{(q)}\}$
0	**24**,48,60,72,**96**
1	**48**,60,84,108,**120**
2	**48**,60,84,108,**120**
3	**60**,72,96,120
4	**60**,72,96,120
5	**60**,72,96,120
6	**60**,72,96,120

of the proliferation rate functions. These nodes have been chosen based upon careful consideration of the data in Figure 3.1 as well as past experience.

Regardless of which parameterization of the proliferation and death rates is used, it should be noted that the current model formulation features proliferation and death rates which are essentially Malthusian in nature (see Section 3.2). That is, the rates at which cells in a particular generation divide and die is assumed to be proportional to the total number of cells in that generation (with 'constants' of proportionality $\alpha_i(t)$ and β_i for proliferation and death, respectively). Alternatively, a model can easily be derived with limiting proliferation and death rates (e.g., logistic rates, Gompertz rates, etc.). Malthusian rates have been used with some success in previous models and should be accurate for any population of cells which divides rapidly enough. Biologically speaking, a cell must proceed through several necessary activities (growth, DNA replication, microtubule formation, etc.) between any two divisions, and this must induce some minimum cell cycle time. Tools such as delay differential equations or stochastic processes have been used to mathematize the cell cycle (see, e.g., [38, 39, 42, 50, 57, 61, 62, 71, 87, 99, 109]) and have resulted in several successful models. We find the current model with its Malthusian rates to be simple and intuitive while also fully capable of accurately fitting the data (Section 3.5). However, it is imperative that the parameters estimated when fitting the model to a particular data set be interpreted in the context of the form of the model being used.

3.4.3 Probabilistically Distributed AutoFI

Up to this point, the derivation and solution of the compartmental model have been shown under the assumption that the natural brightness of cells in the absence of any CFSE molecules, the autofluorescence intensity or AutoFI, can be modeled with sufficient accuracy by a single scalar parameter x_a. However, it is known that the AutoFI of a single cell changes as the cell becomes activated, and that AutoFI varies from cell to cell in the population, even among activated cells.

The AutoFI of cells can be measured directly by setting aside a portion of cells from the PBMC culture which are not labeled with CFSE (but which receive an otherwise identical treatment). The results of such a measurement are shown in Figure 3.6 for two donors, each at two different measurement times. (These data sets were taken independently of the data set shown in Figure 3.1, which is used to calibrate

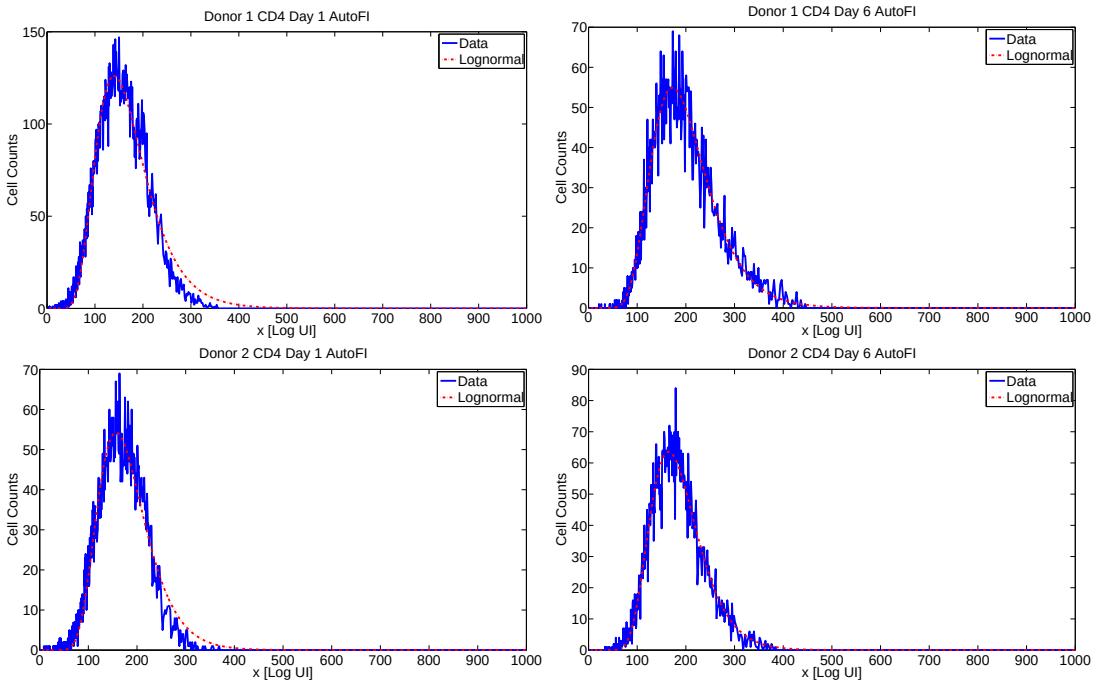

Figure 3.6: Experimentally determined AutoFI distributions with OLS best-fit scaled lognormal curves. Data was taken from two different donors and measured after 24 (left) and 144 (right) hours. We see that a lognormal distribution for AutoFI is quite accurate by $t = 144$ hours (right). Such an assumption is less accurate at $t = 24$ hours, when a significant portion of cells in the population remain unactivated.

the model. Because FI measurements are not absolute–they depend on the calibration and gain settings of the flow cytometer at the time of the experiment–the data shown in Figure 3.6 are intended only to examine the shape of the AutoFI distribution in the population, not its absolute magnitude.) As time progresses, the distribution of AutoFI in the data from both donors increases slightly in mean and is increasingly skewed to the right. These features are also found in additional data sets for $24 < t < 144$ (results unpublished) and appear to be the result of some unmodeled biological processes. The most likely explanation is the known increase in AutoFI as cells become activated [80, Fig. 6]. After a sufficient amount of time, essentially all cells in the culture have either become activated or have died.

Following the discussion at the beginning of Section 3.2, we may consider only AutoFI for activated cells. While we have thus far assumed that this AutoFI can be sufficiently modeled with a single parameter, Figure 3.6 indicates that we might need to consider a probability distribution over the parameter x_a. Let $n(t, x; x_a)$ represent the structured population density of a cohort or subpopulation of cells all of which share the same AutoFI parameter x_a, subject to (3.14). Assume further that this parameter x_a is distributed in the total population of cells with some probability distribution P. Then it follows that the total population is described by

$$\eta(t, x) = E[n(t, x; x_a)|P] = \int_{x_a^{min}}^{x_a^{max}} n(t, x; x_a) dP(x_a). \tag{3.29}$$

It is now clear that the structured density $\eta(t,x)$ for the total population of cells will depend upon the probability measure P. Figure 3.6 depicts the experimental AutoFI data for each donor and measurement time fitted (ordinary least squares) with a scaled lognormal curve. While such an assumption may possibly be of limited validity early in the experiment (probably as a result of the activation process, as discussed above), most cells are undivided at such times and hence the contribution of AutoFI to the total FI of those cells is minimal. Thus we assume that P is reasonably well-described by a lognormal distribution. Hence

$$\frac{dP}{dx_a} = p(x_a) = \frac{1}{x_a\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log x_a - \mu)^2}{2\sigma^2}\right),$$

where

$$\begin{aligned}\mu &= \log(E[x_a]) - \frac{1}{2}\log\left(1 + \frac{Var(x_a)}{E[x_a]^2}\right)\\ \sigma^2 &= \log\left(1 + \frac{Var(x_a)}{E[x_a]^2}\right).\end{aligned}$$

Under such a parametric assumption, the population density $\eta(t,x)$ is uniquely described by the two parameters $E[x_a]$ and $STD[x_a] = \sqrt{Var(x_a)}$ (in addition to the parameters θ discussed so far in this section).

The integral in Equation (3.29) can be easily computed via the midpoint rule. Let $\{x_a^m\}$ be a set of evenly spaced points with spacing Δx_a. Then

$$\eta(t,x) \approx \sum_{m=1}^{M} n(t,x;x_a^m)p(x_a^m)\Delta x_a. \tag{3.30}$$

As written, Equation (3.30) requires the computation of M forward solutions in order to approximate the total population density. However, this computationally intensive approach can be avoided by a change of variables. Define $y = \log_{10}(x - x_a)$ and $\hat{n}(t,y) = 10^y\log(10)n(t,x(y)) = 10^y\log(10)n(t,10^y + x_a)$. Then the system (3.14) becomes

$$\begin{aligned}\frac{\partial\hat{n}_0}{\partial t} - \frac{ce^{-kt}}{\log 10}\frac{\partial\hat{n}_0}{\partial y} &= -(\alpha_0(t) + \beta_0(t) - ce^{-kt})\hat{n}_0(t,x)\\ \frac{\partial\hat{n}_1}{\partial t} - \frac{ce^{-kt}}{\log 10}\frac{\partial\hat{n}_1}{\partial y} &= -(\alpha_1(t) + \beta_1(t) - ce^{-kt})\hat{n}_1(t,x) + 2\alpha_0(t)\hat{n}_0(t,y+\log_{10}2)\\ &\vdots\\ \frac{\partial\hat{n}_{i_{\max}}}{\partial t} - \frac{ce^{-kt}}{\log 10}\frac{\partial\hat{n}_{i_{\max}}}{\partial y} &= -(\beta_{i_{\max}}(t) - ce^{-kt})\hat{n}_{i_{\max}}(t,x) + 2\alpha_{i_{\max}-1}(t)\hat{n}_{i_{\max}-1}(t,y+\log_{10}2).\end{aligned} \tag{3.31}$$

It is plainly observed that the parameter x_a no longer appears in the system of equations for the compartmental model in the structure variable y, while the new initial condition,

$$\hat{\Phi}_0(y) = 10^y\log(10)\Phi_0(10^y + x_a), \tag{3.32}$$

will now depend on x_a. However, provided the initial uptake of CFSE in the experimental procedure

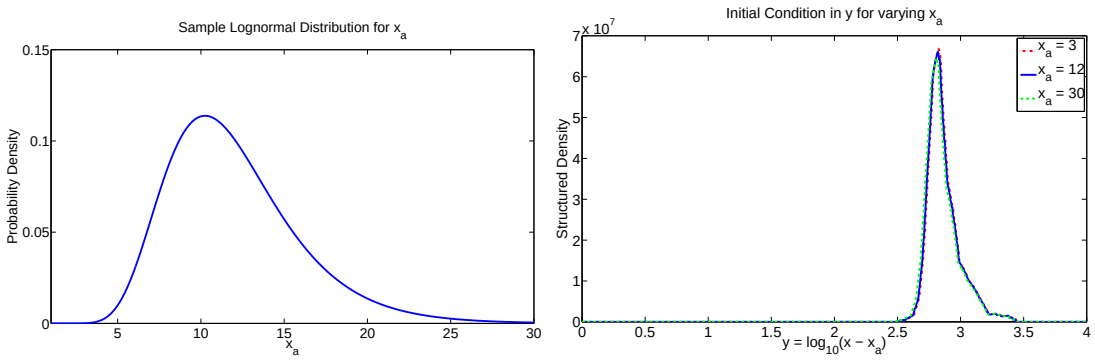

Figure 3.7: Left: A hypothetical lognormal AutoFI distribution with $E[x_a] = 12$ and $Var(x_a) = 4$. Right: Initial conditions (in the structure variable y) computed for the mean value of x_a (solid line) as well as two for two extreme values of x_a. One can see that the value of the parameter x_a has very little effect on the initial condition $\hat{\Phi}_0(y)$.

results in cells with measured FI significantly greater than their AutoFI (which is always the case for useful experimental data), $\Phi_0(x) = 0$ unless $x >> x_a$ (and hence, unless $10^y >> x_a$). As such, the dependence of the initial condition on the parameter x_a can be safely ignored. This fact is demonstrated with an example in Figure 3.7. In general, it is expected that CFSE-labeled cells are approximately 100-1000 times brighter than unlabeled cells (see, e.g., [80, 92, 107]; as mentioned previously, the actual measured FI values depend on machine calibration, and hence will vary from experiment to experiment). Given the initial condition data (Figure 3.3) for our particular data set of interest, it is reasonable to assume $E[x_a] \sim 10$. In Figure 3.7, a sample lognormal distribution with $E[x_a] = 12$ and $STD[x_a] = 4$ is depicted on the left. (These values for the mean and standard deviation can be taken as maximum, worst-case bounds. It is expected that the mean value of x_a is no more than 12, with standard deviation less than 4.) We can assess the effect of the parameter x_a on $\hat{\Phi}_0(y)$ by computing $\hat{\Phi}_0(y)$ for extreme values of x_a (that is, values in the far-left and far-right tails of the density function). The resulting functions (as well as a third function, showing $\hat{\Phi}_0(y)$ when $x_a = E[x_a]$) are shown on the right of Figure 3.7.

It is clear from Figure 3.7 that the initial condition function (for y as a structure variable) changes only minimally for any reasonable values of x_a. (Moreover, the original initial condition $\Phi_0(x)$ was already approximate, having been computed from data in Section 3.3.1.) Thus, computationally, when computing the structured population density according to (3.29), we compute only a single initial condition from Equation (3.32) using $x_a = E[x_a]$. The system (3.31) can then be solved to obtain $\hat{n}(t, y)$ (which does not depend on x_a at all). Next, for each value of x_a in (3.30), one can compute

$$n(t, x; x_a) = \frac{\hat{n}(t, y(x))}{\log(10)(x - x_a)} = \frac{\hat{n}(t, log_{10}(x - x_a)}{\log(10)(x - x_a)},$$

in order to determine the population structured density $\eta(t, x)$. It should be noted that, while the change of variables from x to y eliminates the parameter x_a from the system of PDEs, and we have shown that

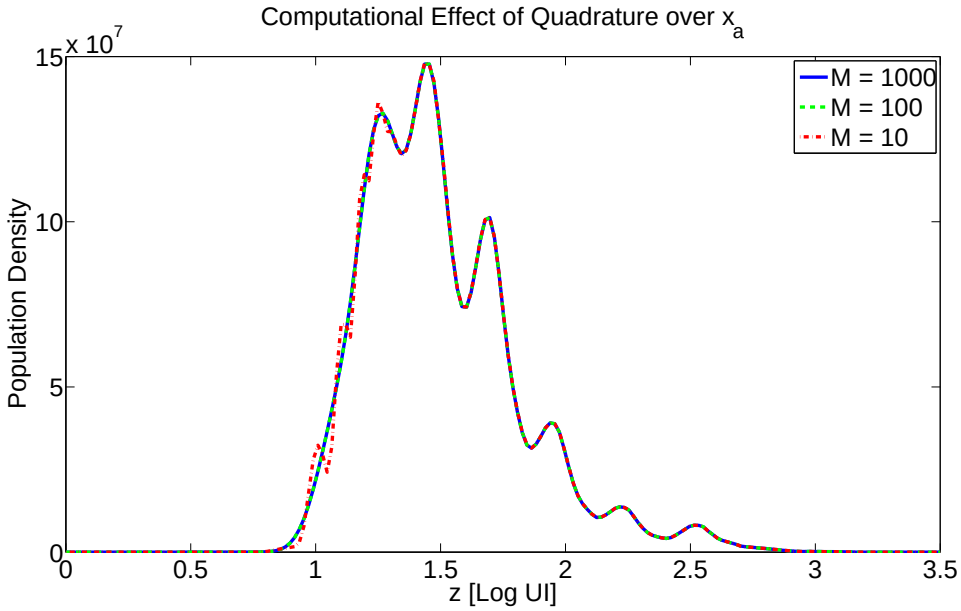

Figure 3.8: Effect of the number of nodes M used to approximate the total population density $\eta(t,x)$ in Equation (3.30). Using $M = 100$ seems more than sufficient.

the effect of x_a on the initial condition $\hat{\Phi}_0(y)$ is negligible, it is not true that the parameter x_a can be ignored entirely. The negligible effect of x_a on the initial condition is the result of the brightness of CFSE-labeled cells at the beginning of the experiment. However, as time progresses, CFSE intensity is lost as cells divide and CFSE degrades, so that AutoFI constitutes a larger percentage of the measured FI. In other words, while it is reasonable to assume $n(0,x) = \Phi_0(x) = 0$ unless $x >> x_a$, this assumption does not hold more generally for $n(t,x)$ $(t > 0)$.

Finally, when using Equation (3.30) to approximate the total population density, one must make sure the parameter M is large enough to provide sufficient accuracy. In Figure 3.8, a sample density is computed at $t = 120$ hours using three different values of M. Given the discussion, above, there is essentially no difference in computational time as M changes. While the solution is not accurately captured for $M = 10$, there is no measurable difference between the solutions for $M = 100$ and $M = 1000$. Henceforth, if it is assumed that AutoFI is distributed in the population of cells, the total population $\eta(t,x)$ will be computed via Equation (3.30) with $M = 100$.

3.4.4 Remarks on the Inverse Problem

At this point, we have considered numerous different parameterizations for the proliferation rate functions $\alpha_i(t)$ (**A1**-**A5**), the death rates β_i (**B1**-**B5**). Each of these parameterizations results in a distinct set of parameters which will need to be estimated from the data. We also have the additional label loss parameters c and k, as well as the AutoFI parameter which can be considered either as a fixed constant x_a or as a lognormal probability distribution with mean $E[x_a]$ and standard deviation $STD[x_a]$.

Table 3.3: Summary of possible parameterizations for the compartmental model, with the set $\theta \in \mathbb{R}^p$ of parameters describing the model in each case.

Model	Parameters	p	Model	Parameters	p
A1B1	$\theta = \{x_a, c, k, \alpha_0, \alpha\}$	5	A1B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha\}$	6
A1B2	$\theta = \{x_a, c, k, \alpha_0, \alpha, \beta\}$	6	A1B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \beta\}$	7
A1B3	$\theta = \{x_a, c, k, \alpha_0, \alpha, \beta_0\}$	6	A1B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \beta_0\}$	7
A1B4	$\theta = \{x_a, c, k, \alpha_0, \alpha, \beta_0, \beta\}$	7	A1B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \beta_0, \beta\}$	8
A1B5	$\theta = \{x_a, c, k, \alpha_0, \alpha, \{\beta_i\}\}$	12	A1B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \{\beta_i\}\}$	13
A2B1	$\theta = \{x_a, c, k, \{\alpha_i\}\}$	9	A2B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}\}$	10
A2B2	$\theta = \{x_a, c, k, \{\alpha_i\}, \beta\}$	10	A2B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \beta\}$	11
A2B3	$\theta = \{x_a, c, k, \{\alpha_i\}, \beta_0\}$	10	A2B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \beta_0\}$	11
A2B4	$\theta = \{x_a, c, k, \{\alpha_i\}, \beta_0, \beta\}$	11	A2B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \beta_0, \beta\}$	12
A2B5	$\theta = \{x_a, c, k, \{\alpha_i\}, \{\beta_i\}\}$	16	A2B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \{\beta_i\}\}$	17
A3B1	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha\}$	6	A3B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha\}$	7
A3B2	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \beta\}$	7	A3B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \beta\}$	8
A3B3	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \beta_0\}$	7	A3B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \beta_0\}$	8
A3B4	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \beta_0, \beta\}$	8	A3B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \beta_0, \beta\}$	9
A3B5	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \{\beta_i\}\}$	13	A3B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \{\beta_i\}\}$	14
A4B1	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}\}$	10	A4B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}\}$	11
A4B2	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \beta\}$	11	A4B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \beta\}$	12
A4B3	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \beta_0\}$	11	A4B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \beta_0\}$	12
A4B4	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \beta_0, \beta\}$	12	A4B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \beta_0, \beta\}$	13
A4B5	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \{\beta_i\}\}$	17	A4B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \{\beta_i\}\}$	18
A5B1	$\theta = \{x_a, c, k, \{a_i^{(p)}\}\}$	21	A5B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}\}$	22
A5B2	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \beta\}$	22	A5B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \beta\}$	23
A5B3	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \beta_0\}$	22	A5B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \beta_0\}$	23
A5B4	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \beta_0, \beta\}$	23	A5B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \beta_0, \beta\}$	24
A5B5	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \{\beta_i\}\}$	28	A5B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \{\beta_i\}\}$	29

In the remainder of this report, we will refer to the model solution simply as $n(t, x; \theta)$ where $\theta \subset \mathbb{R}^p$ is a set of parameters which describes the model. (This includes the case that x_a is described by a probability measure, where $\eta(t, x)$ was used in the previous exposition.) This is done to simplify notation, and it will always be clear from context which parameterization is being used. Obviously, the value of p will vary depending upon the parameterization. The various possibilities are summarized in Table 3.3.

We now return to the OLS formulation (3.27) of the inverse problem,

$$\theta_{OLS} = \arg\min_{\theta \in \Theta} \sum_{k,j} \mathcal{R}_{kj}^2 = \arg\min_{\theta \in \Theta} \sum_{k,j} \left(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j \right)^2,$$

where now Θ is a closed bounded subset of $\mathbb{R}^p$. Using the data $\{n_k^j\}$ as realizations of the random variables $\{N_k^j\}$, we would like to compute the estimate

$$\hat{\theta}_{OLS}(n_k^j) = \arg\min_{\theta \in \Theta} J(\theta | n_k^j) = \arg\min_{\theta \in \Theta} \sum_{k,j} \left(I[\tilde{n}](t_j, z_k^j; \theta) - n_k^j \right)^2. \tag{3.33}$$

However, we have only an approximate numerical solution with which to compare the data. Thus we

Table 3.4: Summary of bound-constraints for the OLS parameter estimation problem (3.34). The parameters $\{\alpha_i\}$, $\{a_i^{(p)}\}$, and $\{\beta_i\}$ must be positive. The feasibility of the remaining bounds has been determined computationally.

Parameter	Minimum	Maximum
x_a	1	20
$E[x_a]$	5	12
$STD[x_a]$	0	4
c	1×10^{-4}	1×10^{-2}
k	0	1×10^{-3}
$\{\alpha_i\}$ or $\{a_i^{(p)}\}$	0	1
$\{\beta_i\}$	0	1

actually compute the approximate estimate

$$\hat{\theta}_{OLS}(h_t, N_x, M; n_k^j) = \arg\min_{\theta\in\Theta} J_A(\theta|n_k^j) = \arg\min_{\theta\in\Theta} \sum_{k,j} \left(I_A[\tilde{n}](t_j, z_k^j; \theta) - n_k^j\right)^2, \tag{3.34}$$

where we have now explicitly emphasized the dependence of the parameter estimate on the computational accuracy of the numerical solution. The continuous dependence of the model solution $\tilde{n}(t, z; \theta)$ on the parameter θ (regardless of which particular parameterization is used) follows easily from the method of characteristics solution of Section 3.3. Numerical convergence with respect to h_t, N_x, and M follow directly from well-known results regarding the trapezoidal rule for quadrature, linear interpolation of a smooth function, and the midpoint rule for quadrature, respectively. As such, it can be shown (see, e.g., the arguments of [16, Ch. 3] that the approximate estimates $\hat{\theta}_{OLS}(h_t, N_x, M; n_k^j)$ will converge to some $\hat{\theta}_{OLS}^*$ which minimizes (3.33) as $N_x, M \to \infty$, and $h_t \to 0$. It should be noted that the possible nonuniqueness of the minimizer $\hat{\theta}_{OLS}^*$ is a common issue in inverse problems. We forgo techniques such as Tikhonov regularization in this report, choosing to focus instead on the accuracy of the best fit models $n(t, z; \hat{\theta}_{OLS})$ in fitting a particular data set, regardless of uniqueness (although these issues must be dealt with in order to establish standard errors, confidence intervals, etc.). For the remainder of this report, we will not distinguish between $\hat{\theta}_{OLS}$, $\hat{\theta}_{OLS}^*$, or $\hat{\theta}_{OLS}(h_t, N_x, M; n_k^j)$. It should also be noted that this best-fit parameter, which is itself an estimate of the random variable θ_{OLS}, will be data-realization dependent. However, for a good model and a sufficiently large data set, $\hat{\theta}_{OLS}$ is an unbiased estimator of θ_{OLS} [22, 37, 96].

The optimization (3.34) has been implemented in MATLAB using the `fmincon` function, which is a variation of the BFGS-active set algorithm for bound-constrained parameters. The parameter constraints are summarized in Table 3.4.

3.4.5 Information Theoretic Model Selection

Each possible parameterization presented thus far gives rise to a distinct mathematical model which can be fit to a data set in the prescribed manner. Based upon the results for each model, we would like to determine which parameterization is most appropriate and use those results to draw conclusions

regarding division-linked and/or longitudinal changes in the behavior of the cell culture. In order to do this, we must establish some formal mechanism which permits the objective comparison of different models.

One common approach is hypothesis testing for model refinements [8, 10]. However, such methods are only useful for pairwise comparisons, and are better suited for comparison against an experimental control [31]. Moreover, such methods do not apply unless one of the two models in the comparison is contained within the other model (for instance, our parameterization **B4** contains **B3** as a special case). While several parameterizations discussed in this document are indeed contained within other parameterizations, this is not universally the case (e.g., there is no containment between **B2** and **B3**).

A more general approach, based upon the premises of information theory, is found in the Akaike Information Criterion (AIC). Briefly, for models with independent, homoscedastic, normally distributed errors, it can be shown that

$$AIC = m \log\left(\frac{J(\hat{\theta}_{OLS})}{m}\right) + 2p, \tag{3.35}$$

where m is the total number of data points, is an approximately unbiased estimate of the "expected relative Kullback-Leibler distance" (information loss) when a model is used to describe a data set [31]. Given a set of R models, the AIC can be computed for each model; we seek the model which results in the smallest AIC value. It should be emphasized that the AIC is only an estimate of information loss, and this estimate depends on the particular data set being used. (When comparing different models with the AIC, the same data set must be used to fit each model.) As discussed in Section 3.4.1, we cannot ascertain a priori that the measurement errors are normally distributed with constant variance. However, the use of an OLS framework already constitutes an assumption of homoscedasticity. The assumption of normality does not seem to be a significantly greater burden, and we proceed with the AIC in spite of these issues. As will be shown in Section 3.5, there is a very clear preference among the models when ranked by the AIC. As such, we do not expect our results to change significantly for a different error model. Similarly, the derivation of the AIC assumes that any model which is fit to the data is sufficiently accurate so that the assumption $E[N_k^j] = n(t_j, x_j; \hat{\theta}_{OLS})$ is valid. While this assumption may break down for the least accurate of the models tested in this report (see Section 3.5), it is a standard assumption in the OLS framework (3.27), provided the estimate $\hat{\theta}_{OLS}$ is sufficiently close to θ_0 for each particular model. The AIC is motivated at slightly greater length in Chapter 5.

There is an element of parsimony in the AIC, as a model which fits the data poorly (high $J(\hat{\theta}_{OLS})$) or which contains a large number of parameters (high p) will have a comparatively larger AIC. Yet, rather than using the AIC to determine a single 'best' model, additional theory is available. If AIC_{min} is the smallest computed AIC value, then we can define the AIC differences

$$\Delta_r = AIC_r - AIC_{min}, \tag{3.36}$$

for $1 \leq r \leq R$, where AIC_r is the AIC value computed when model r is fit to the data. Finally, we can compute the Akaike weights

$$w_r = \frac{\exp\left(\frac{-\Delta_r}{2}\right)}{\sum_r \exp\left(\frac{-\Delta_r}{2}\right)}. \tag{3.37}$$

Table 3.5: Summary of results for the various models considered in this report. The AIC-selected best model, parameterization A5B5 with lognormally distributed AutoFI, not only has the lowest cost, the OLS cost of this model is so much smaller (compared to the other models tested) that its AIC value is *significantly* lower than for any other model. The Akaike weights are not shown, as the weight assigned to model A5B5dist must be greater than $1 - 50\exp(-43/2) > 1 - 1 \times 10^{-7}$.

Model	$J_A(\hat{\theta}_{OLS})$	AIC_r	Δ_r	Rank	Model	$J_A(\hat{\theta}_{OLS})$	AIC_r	Δ_r	Rank
A1B1	48.9309×10^{11}	89459	11850	50	A1B1dist	44.5493×10^{11}	89059	11450	45
A1B2	48.5765×10^{11}	89430	11821	49	A1B2dist	26.0134×10^{11}	86753	9144	31
A1B3	46.0968×10^{11}	89205	11596	48	A1B3dist	18.9754×10^{11}	85400	7791	26
A1B4	46.0439×10^{11}	89202	11593	47	A1B4dist	18.9754×10^{11}	85402	7793	27
A1B5	35.5868×10^{11}	88107	10498	40	A1B5dist	17.9868×10^{11}	85183	7574	24
A2B1	30.8384×10^{11}	87487	9878	37	A2B1dist	42.3075×10^{11}	88845	11236	43
A2B2	30.4566×10^{11}	87436	9827	36	A2B2dist	21.2095×10^{11}	85886	8277	29
A2B3	28.6677×10^{11}	87176	9567	33	A2B3dist	14.8563×10^{11}	84359	6750	16
A2B4	28.6677×10^{11}	87178	9569	34	A2B4dist	14.8562×10^{11}	84361	6752	17
A2B5	28.6677×10^{11}	87188	9579	35	A2B5dist	14.8562×10^{11}	84371	6762	18
A3B1	45.4019×10^{11}	89140	11531	46	A3B1dist	42.9086×10^{11}	88900	11291	44
A3B2	37.3875×10^{11}	88309	10700	42	A3B2dist	11.9759×10^{11}	83428	5819	11
A3B3	34.8434×10^{11}	88007	10398	38	A3B3dist	13.5090×10^{11}	83945	6336	12
A3B4	34.8376×10^{11}	88008	10399	39	A3B4dist	10.5215×10^{11}	82875	5266	10
A3B5	18.7334×10^{11}	85357	7748	25	A3B5dist	6.9142×10^{11}	81084	3475	8
A4B1	25.3453×10^{11}	86648	9039	30	A4B1dist	36.6830×10^{11}	88236	10627	41
A4B2	16.8159×10^{11}	84890	7281	22	A4B2dist	5.5690×10^{11}	80152	2543	7
A4B3	17.1422×10^{11}	84973	7364	23	A4B3dist	8.3562×10^{11}	81893	4284	9
A4B4	16.6846×10^{11}	84859	7250	21	A4B4dist	5.0699×10^{11}	79752	2143	6
A4B5	16.4652×10^{11}	84812	7203	20	A4B5dist	4.5712×10^{11}	79318	1709	5
A5B1	19.6228×10^{11}	85572	7963	28	A5B1dist	27.3143×10^{11}	86993	9384	32
A5B2	15.0638×10^{11}	84440	6831	19	A5B2dist	3.5086×10^{11}	78193	584	4
A5B3	14.6710×10^{11}	84327	6718	13	A5B3dist	3.2607×10^{11}	77879	270	3
A5B4	14.6740×10^{11}	84330	6721	14	A5B4dist	3.0918×10^{11}	77653	43	2
A5B5	14.6727×10^{11}	84339	6730	15	A5B5dist	3.0535×10^{11}	77609	0	1

It can be shown (either by likelihood ratio tests or in a Bayesian framework, see [31]) that the AIC weight w_r can be interpreted as the probability that model r is the best model to describe the data (given the set of R possible models). Thus, after each model from Table 3.3 is fit to a data set, we can compute the Akaike weights for the set of candidate models and use these to assess the necessity of various mathematical features (e.g., division dependence of cell death rates) in describing the data. A complete derivation of the AIC and Akaike weights, as well as numerous examples and exhaustive references, can be found in [31].

3.5 Results and Discussion

The model calibration results for each possible parameterization of the compartmental model considered in this report are summarized in Table 3.5. The approximate OLS costs $J_A(\hat{\theta}_{OLS})$ are shown for each parameterization, as well as the computed AIC values and AIC differences. The models are also ranked in terms of their relative information theoretic loss.

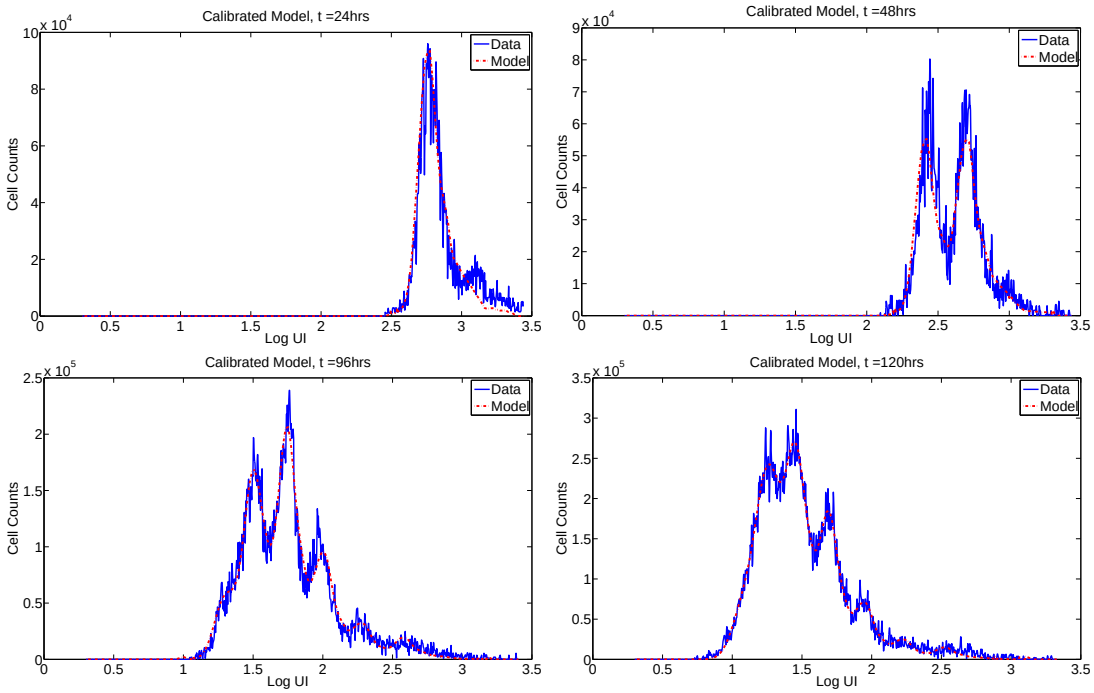

Figure 3.9: Best-fit solution $I_A[\tilde{n}](t, z; \hat{\theta}_{OLS})$ for parameterization A5B5dist. Total cost $J_A(\hat{\theta}_{OLS}) = 3.0535 \times 10^{11}$.

The AIC selected model is parameterization A5B5 with lognormally distributed AutoFI (henceforth, A5B5dist) with a cost $J_A(\hat{\theta}_{OLS}) = 3.0535 \times 10^{11}$. This parameterization resulted in a model with not only the smallest AIC value, but also the lowest cost (meaning that the decrease in cost more than offset the additional parameters). The optimal solution for parameterization A5B5dist is depicted in comparison to the data in Figure 3.9. The estimated piecewise linear proliferation rates can be found in Figure 3.10, and the estimated death rates are summarized in Table 3.6. For the AutoFI distribution, the best-fit lognormal distribution has mean $E[x_a] = 8.739$ UI and $STD[x_a] = 3.534$ UI. The estimated Gompertz label decay parameters are $c = 5.641 \times 10^{-3}$ and $k = 1 \times 10^{-9}$.

As can clearly be seen in Figure 3.9, the compartmental model (with suitable parameterization) is capable of accurately describing the particular data set used for model calibration in this report. The most notable shortcoming of the model occurs at $t = 24$ hours, where a distinct cohort of cells with high CFSE FI can be seen in the data and is not modeled accurately. As discussed in the introduction, this cohort is believed to be either cell duplets or some other anomalous cell types which were not properly gated out of the measured cell data, and such cells should not be an issue in future data sets. It also appears that neither of the two generations in the model solution at $t = 48$ hours contains enough cells (when compared to the data at that time). This may also be partly explained as a systematic error resulting from the presence of cell duplets in the data. It is also possible that small errors associated with the manner in which counted beads (see Section 1.2) are used to determine the total population size.

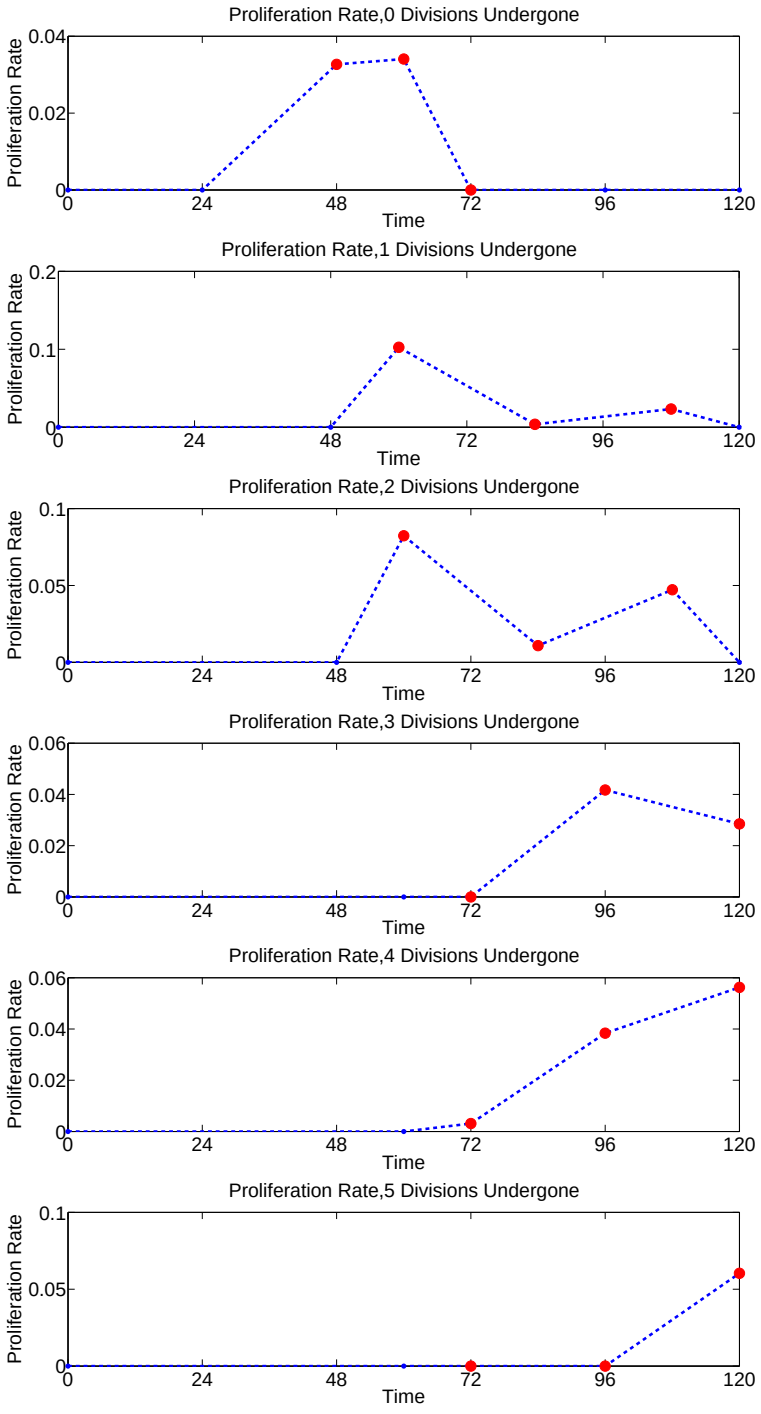

Figure 3.10: OLS best-fit piecewise linear proliferation rate functions for each division number. Red circles indicate nodes which were estimated in the inverse problem.

Table 3.6: Estimated death rates β_i in terms of the number i of divisions undergone.

Divisions	Death Rate (1/hr)
0	0.0165
1	0.0000
2	0.0000
3	0.0000
4	0.0012
5	0.0544
6	0.1572

One of the primary goals of considering various parameterizations for the proliferation and death rates (Section 3.4) was to investigate the dependence of these rates on division number and on time. The best-fit parameterization A5B5dist features a proliferation rate which depends both on time and division number, as well as a death rate which depends on division number. Additionally, the AutoFI parameter x_a is lognormally distributed, which was not considered in previous efforts [20]. Given the overwhelming weight assigned to the parameterization A5B5dist in the information theoretic framework, it is tempting to conclude that each of these features is necessary in accurately modeling the data. Because the data set contains 4289 points, even a small difference in OLS cost (compare, for example, parameterizations A5B5dist and A5B4dist) results in significantly different AIC values. However, the AIC (3.35) is derived under the assumptions of independent, homoscedastic, normally distributed errors. If these assumptions break down, particularly if the 4289 points are not independent, then the magnitude of the AIC differences may be misleadingly large.

In spite of these potential setbacks, there are still several useful conclusions which can be safely drawn. As expected, the worst parameterizations (in terms of both OLS cost and AIC rank) are those which do not permit cell death in the population (**B1**). Parameterizations which feature probabilistically distributed AutoFI are more accurate than parameterizations which use a constant parameter x_a to describe AutoFI. Among the models which use a constant parameter x_a to describe AutoFI, the most parsimonious model (that is, the AIC selected model) is parameterization A5B4. (Parameterizations A5B4 and A5B5 differ minimally in cost, but A5B4 has fewer parameters.) The best-fit solution for this model is shown in comparison to the data in Figure 3.11. *We find that a model which fails to account for variability in AutoFI in the population of cells does not adequately describe the increasing heterogeneity of the population of cells as division number increases.* This is particularly noteworthy for cells having undergone 4 or more divisions, where AutoFI constitutes a comparatively larger fraction of the measured FI of the cells. Such an observation has important experimental ramifications for the design of intracellular dyes. While it has long been known that a population of cells must obtain a high level of FI (relative to their AutoFI) during the initial staining process in order for the experimenter to resolve multiple rounds of division in the population [80, 92], we now see that the variability of AutoFI in the population of cells also has an effect on the peak-to-peak resolution of the data. While AutoFI is a property of the cells being measured (it arises from intracellular molecules which emit light in the frequency bands used to detect the intracellular dye), *focus may possibly be directed toward the design of dyes with spectral properties that minimally overlap with common intracellular molecules.*

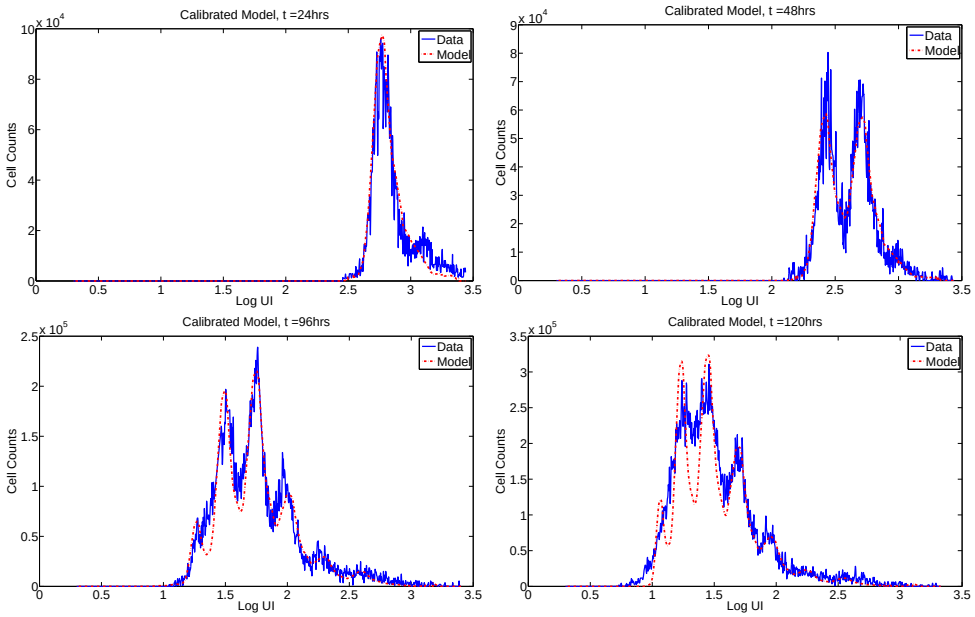

Figure 3.11: Among the models which do not use a lognormal distribution to describe AutoFI, the AIC selected model is parameterization A5B4. When comparing the best-fit solution to the data, it is clear that a model lacking an AutoFI distribution will result in peaks which are too distinct when compared to the data.

As in previous work [20, 21], we find that time dependence is a significant feature of the proliferation rates, given the model formulation (3.14). Significantly, we find that the population of cells cannot be accurately modeled by considering only a delay in the time to first division. For instance, the calibrated model using parameterization A4B5dist (which is the AIC selected model among those which does not feature completely time dependent proliferation) is shown in Figure 3.12. This parameterization does not permit any proliferation until $t \geq t^*$, thus enforcing a delay before any division occurs in the population. Even with this feature, subsequent divisions of cells emerge too quickly in the model solution. Thus more complex time-dependence (parameterization **A5**) appears to be necessary, as the resulting decrease in cost outweighs the additional parameters.

The compartmental model was motivated by a desire to compute quantities such as cell numbers from the best-fit model solution. As discussed in Chapter 1, previous methods for obtaining cell numbers relied on some form of deconvolution of the histogram data, typically via fitting by a series of normal or lognormal curves. While the compartmental model is more mathematically involved and requires considerably more time for fitting to data (a few minutes to a few hours, depending upon the parameterization used and the accuracy of the initial parameter guess for the BFGS algorithm), it does not require any assumption as to the shape of the distribution of cells within a single generation. Given a calibrated model solution $n(t, x; \hat{\theta}_{OLS})$, one can compute the total number of cells

$$N_i(t) = \int_{x_a}^{\infty} n_i(t, x; \hat{\theta}_{OLS}) \tag{3.38}$$

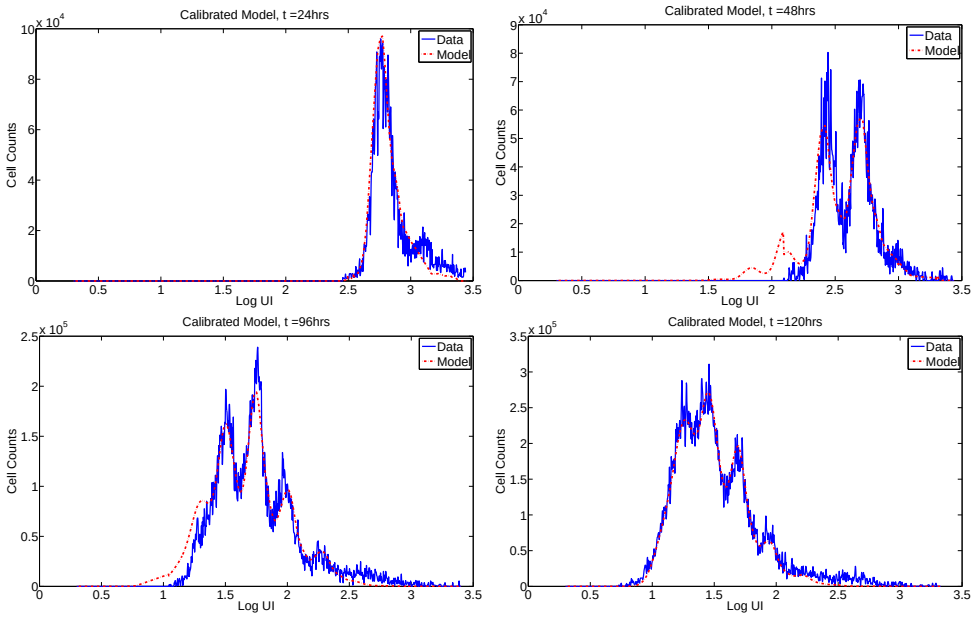

Figure 3.12: Best-fit model solution with parameterization A4B5dist, the AIC selected model among the subset of models which does not feature completely time-dependent proliferation. While this parameterization includes a delay before the first division is reached, this is still insufficient to describe the data as cells proceed through subsequent rounds of division too quickly. The discontinuity in the model solution at $t = 48$ hours is a result of a sudden change in the size of the histogram bins on which the data is specified.

for each generation. It may also be of experimental interest to consider the number of precursors in the population. Because each cell division results in the formation of two daughter cells from a single mother cell, one must renormalize (by a factor of 2) the total number of cells in each generation in order to accurately analyze the proportion of cells proceeding through a specified number of divisions. Precursors, then, are cells in the original population (that is, at $t = 0$ hours) which eventually give rise to other cells with higher division numbers at later times. The number of precursors is

$$P_i(t) = \frac{N_i(t)}{2^i} = \frac{1}{2^i}\int_{x_a}^{\infty} n_i(t, x; \hat{\theta}_{OLS}). \tag{3.39}$$

Given that precursors represent numbers of cells in the original population, it follows that the total number of precursors

$$P(t) = \sum_{i=0}^{i_{\max}} P_i(t)$$

cannot increase in time (but may decrease as a result of cell death). Cell numbers and precursor numbers have been computed from the best-fit model solution (parameterization A5B5dist) and are shown in Figure 3.13 for undivided cells ($N_0(t)$, $P_0(t)$) and divided cells ($\sum_{i=1}^{i_{\max}} N_i(t)$, $\sum_{i=1}^{i_{\max}} P_i(t)$) as well as total cells. It follows that such curves could easily be used to determine such parameters as approximate

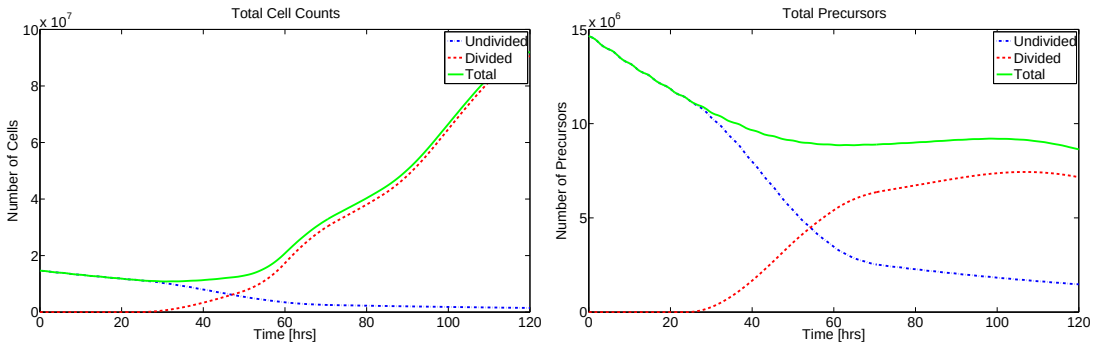

Figure 3.13: Total cell counts (left) and total precursors (right) in terms of undivided cells, divided cells, and total cells. The values are computed from the best-fit model solution $n(t, x; \hat{\theta}_{OLS})$ with parameterization A5B5dist. Numerical values at data collection times are summarized in Tables 3.7 and 3.8. The slight increase (less than 0.3%) in the total number of precursors between $t \approx 60$ hours and $t \approx 90$ hours is within the range of numerical error for the computed solution.

doubling time for the population, or the fraction of cells which do not divide. These parameters may be of particular importance in accounting for changes in behavior as experimental conditions (e.g., strength of stimulation) or in a diagnostic setting.

3.6 Discussion

In this chapter, a label structured system of PDEs for a population of dividing cells, indexed by the number of generations undergone, is derived and fit to data. Under the appropriate assumptions for the label loss rate, autofluorescence parameter, proliferation rates, and death rates, such a model can accurately fit an experimental data set (Figure 3.9). Because each generation of cells is mathematically described by a separate structured density function, the proliferation and death rates can be estimated directly in terms of division number, and there is no need for any parameterization of these rates in the structure variable. This is in contrast to the previous fragmentation model (3.1) from Chapter 2. The AIC-selected best-fit compartmental model contains 29 parameters and results in a best-fit OLS cost of $J_A(\hat{\theta}_{OLS}) = 3.0535 \times 10^{11}$ while the best-fit fragmentation model contained 73 parameters and resulted in a cost of 3.0901×10^{11}. Thus the compartmental model appears quite superior to the previous fragmentation model, as it contains fewer parameters and has a lower OLS cost. Additionally, the compartmental model can be used to compute cell numbers in terms of the number of divisions undergone. Certainly it may be possible to decrease the number of parameters in the fragmentation model by changing the placement of the nodes used for the estimation of the proliferation and death rate functions. Yet even if the total number of parameters could be decreased significantly without increasing the OLS cost of the fragmentation model, it still could not be used directly to compute cell numbers.

Table 3.7: Total numbers of cells in terms of division number. For each time and generation, the total number of cells has been computed from the OLS best-fit model solution (top), from a deconvolution of the data using normal curves (middle) and from a deconvolution of the data using lognormal curves (bottom). While the numbers computed from normal and lognormal curves are generally close together, there are clear differences between the values computed from the deconvolution methods and those obtained with the compartmental model. The most striking example occurs for $t = 120$ hours for cells having undergone 6 divisions. It is interesting to note that the division peaks in the histogram data are not well-resolved for such cells, making the accurate determination of cell numbers difficult.

	Divisions Undergone							
Time (hrs)	0	1	2	3	4	5	6	Total
	11339892	0	0	0	0	0	0	11339892
24	9211571	0	0	0	0	0	0	9211571
	9254681	0	0	0	0	0	0	9254681
	5881557	6555814	0	0	0	0	0	12437372
48	6298359	5473128	0	0	0	0	0	11771487
	6294945	5434570	0	0	0	0	0	11729515
	1906065	2478042	8092563	20976431	18520420	7588997	0	59562519
96	1970401	3284000	11019352	18184100	18307586	5252346	0	58017785
	2364520	3940800	10467773	17773515	18846649	5406993	0	58800249
	1476266	1978930	5605926	17086869	25529315	25986246	14337080	92000632
120	1969773	2969295	7881600	21017600	26272000	24958400	4597599	89666268
	2195762	3435673	7722653	17150087	26755781	29950079	5517118	92727154

3.6.1 Comparison to Deconvolution Techniques

Given the applicability of the compartmental model (once calibrated) to computing the numbers of cells having undergone a specified number of divisions, a relevant comparison can be drawn between the results of a label structured PDE model and the commonly used deconvolution techniques. In Table 3.7, the number of cells in each generation is computed at each measurement time. For each time and generation, the top number is computed from Equation (3.38). The middle number is computed by first fitting the function

$$\psi(z_k^j) = \sum_{i=0}^{i_{\max}} \psi_i(z_k; s_i, \mu_i, \sigma_i)$$

to the data at a given time, where $\psi_i(z_k; k_i, \mu_i, \sigma_i)$ is a normal density function with mean μ_i and standard deviation σ_i, scaled by a factor s_i. The method of fitting is ordinary least squares. Then the total number of cells having undergone i divisions is $\sum_k \psi_i(z_k; s_i, \mu_i, \sigma_i)$. The final number in each block of Table 3.7 is computed in an analogous manner, but with lognormal rather than normal density functions.

Unsurprisingly, the two deconvolution techniques (fitting with a series of normal or lognormal curves) provide estimates of cell numbers which are fairly consistent. However, these estimates occasionally differ from estimates obtained from the compartmental model. Of particular note is the difference for cells having undergone 6 divisions at $t = 120$ hours. It should be noted that this generation of cells is quite hard to distinguish in this particular histogram data set. Such poorly resolved generations of cells can be quite problematic for the deconvolution techniques, as the unique estimation of parameters (for the

Table 3.8: Total precursors in terms of divisions number. For each time and generation, the total number of precursors has been computed from the OLS best-fit model solution (top), from a deconvolution of the data using normal curves (middle) and from a deconvolution of the data using lognormal curves (bottom). As in Table 3.7 we find general agreement between the values computed from deconvolution techniques, which are slightly different than those computed from the compartmental model. Under the assumptions of the experiment, the total number of precursors should not increase.

	Divisions Undergone							
Time (hrs)	0	1	2	3	4	5	6	Total
24	11339892	0	0	0	0	0	0	11339892
	9211571	0	0	0	0	0	0	9211571
	9254681	0	0	0	0	0	0	9254681
48	5881557	3277907	0	0	0	0	0	9159465
	6298359	2736564	0	0	0	0	0	9034923
	6294945	2717285	0	0	0	0	0	9012230
96	1906065	1239021	2023141	2622054	1157526	237156	0	9184963
	1970401	1642000	2754838	2273013	1144224	164136	0	9948612
	2364520	1970400	2616943	2221689	1177916	168969	0	10520436
120	1476266	989465	1401482	2135859	1595582	812070	224017	8634740
	1969773	1484648	1970400	2627200	1642000	779950	71837	10545808
	2195762	1717837	1930663	2143761	1672236	935940	86205	10682405

normal or lognormal curves) requires that distinct generations of cells be plainly visible. It appears to be a major advantage of the compartmental model to be able to fit data (and hence compute cell numbers) even when the histogram data features generations of cells which are less than ideally resolved. Of course, it is not possible to say from these results which technique (if either) is providing the correct number of cells. Yet, because the compartmental model is derived from a conservation law, and this conservation law must hold regardless of the parameters input into the model, cells cannot enter or leave the population except as permitted by the form of the model and the given parameters. Meanwhile, the deconvolution techniques do not arise from any conservation law, and the computed cell numbers in each generation may increase or decrease freely, unrestrained by any balance law. It seems then, that the compartmental model should have a major advantage in computing cell numbers, owing to its 'memory' of the number of cells determined at previous time points (even when the generations of cells are poorly resolved in the data).

This is particularly noteworthy in Table 3.8, where the number of precursors for each generation of cells is computed. As in Table 3.7, each block of Table 3.8 contains the results computed from the compartmental model, deconvolution with normal curves, and deconvolution with lognormal curves. Observe the significant increase (more than 10%) in the total number of precursors as computed by deconvolution between $t = 48$ and $t = 96$ hours. As discussed above, the total number of precursors cannot increase in a population of cells. While the number of precursors computed from the compartmental model also increases, it does so by a very small amount (less than 0.3%) consistent with the error in the numerical solver. Of course, it has already been noted that some data sets do in fact exhibit increases in the total number of precursors–a discrepancy arising from the inaccuracy of the assumption that each well plate contains an identical population of cells. On one hand, the deconvolution techniques would seem to have an advantage, as they are not constrained by any conservation law. However, this has an interesting

implication. Because the deconvolution techniques do not link the population estimates from one data collection time to the next, there is a potential bias associated with such methods as a result of sample-to-sample variability in the experimental data. It should be noted that sample-to-sample variability is also problematic for the compartmental model solution. If the samples used to obtain the experimental data are not sufficiently similar, the conservation law (which follows from the assumption that each sample is identical) used to derive the model may not hold. In such a case, the compartmental model would be systematically in error (when compared to the data), as the calibrated model itself would still follow the assumed conservation law. Following the discussion above, we believe that a more accurate statistical model, which will necessarily include a careful consideration of the method of sampling/data collection, will resolve any discrepancy with the compartmental model. Some preliminary work on this subject is surveyed in the next two chapters.

3.6.2 Concluding Remarks

It is interesting to note that using the compartmental model we have found variability in AutoFI to be an essential feature of an accurate mathematical model. Yet the fragmentation model assumes only a constant value of AutoFI without significant sacrifice in accurately fitting the data (compare, e.g., Figures 2.12 and 3.11). The explanation for this unusual observation is the manner in which the proliferation rate is parameterized as a function of the structure variable (or the 'translated variable') in the fragmentation model. The large number of nodes used for the structural dependence of that proliferation rate (13 nodes) allows for significant variability in the proliferation rate, even among cells which are sufficiently close in the structural coordinate. Because the Gompertz decay function for label decay assumes that the rate of FI loss is directly proportional to the quantity of FI, a group of cells which divides immediately and then pauses will lose less label than a group of cells which waits for some time and then divides. In other words, the variability of the proliferation rate induces a variability in the label loss rate. As a consequence of this observation, it would be interesting to compare the effects of probabilistically distributed AutoFI with the effects of probabilistically distributed label loss rates in the compartmental model.

The major advantage of the compartmental model over previous efforts is the ability to compute cell numbers directly from the model solution (Figure 3.13). Because the compartmental model can be used to estimate the numbers of cells (or precursors) having undergone a specified number of divisions, biologically meaningful parameters can be assessed directly in terms of division number. For instance, the total number of precursors in the population, as a fraction of the original number, provides a meaningful estimation of cell viability. The total number of cells in the population can be used to estimate the population doubling time. As more complex experiments are conducted, the compartmental model could be easily generalized to account for division-linked changes (surface marker expression/differentiation, genetic mutations, etc.). Such features should be useful when comparing results from different data sets, such as when attempting to quantify the effects of a given chemical reagent, or distinguishing between diseased and healthy cells. Additional generalizations are discussed at greater length in Chapter 5.

Of course, the meaningful comparison of parameter estimates between multiple data sets and experimental conditions relies upon quantification of the levels of uncertainty in the estimated parameters.

This quantification, typically in the form of confidence bounds, is premised upon the accurate specification of the statistical model (3.25) which links the model to the data. In this chapter, the model was fit to the data in an ordinary least squares sense, with the tacit assumption that the error random variables $\mathcal{E}_{kj}$ have mean zero and constant variance. However, as will be shown in the next chapter, this assumption is not an accurate description of the data. While the misspecification of the statistical error model does not invalidate the ability of the compartmental model to (qualitatively) fit the available CFSE data set, it does prevent the meaningful quantification of uncertainty in the parameter estimates.

In an effort to form a more reliable statistical model, in the next chapter we begin an analysis of the OLS residuals after the AIC-selected best-fit model (parameterization A5B5dist) is fit to the data (Figure 3.9). This is followed by a qualitative analysis of several additional data sets which were recently collected for the purposes of analyzing variability in the data collection procedure. These analyses are then used to suggest potential improvements to the statistical model.

Chapter 4

The Statistical Model and Data Variability

4.1 Motivation and Goals

At this point, a mathematical model (3.14) has been derived which can be fit accurately and directly to a CFSE data set. Using this model, it is possible to examine several biologically relevant measures of a proliferative response, such as rates of cell division and death, cell viability, and population doubling time. It is hoped that such a model will provide a quantitative framework for the comparison of data sets arising from cells in various biological and experimental conditions. However, before such a framework can be established, there is a need for meaningful confidence intervals to quantify the certainty with which individual parameters are estimated. This, in turn, relies upon an accurate statistical model for the CFSE histogram data.

Recall from Section 3.4.1 the assumption that the data is accurately described by the statistical model

$$N_k^j = I[\tilde{n}](t_j, z_k^j; \theta_0) + \mathcal{E}_{kj}, \tag{4.1}$$

where $\mathcal{E}_{kj}$ are independent random variables satisfying $E[\mathcal{E}_{kj}] = 0$. Then the best-fit parameter estimate is

$$\hat{\theta}_{WLS}(n_k^j) = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{r_{kj}^2}{w_{kj}} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{1}{w_{kj}} \big(I[\tilde{n}](t_j, z_k^j; \theta) - n_k^j\big)^2, \tag{4.2}$$

where the residuals r_{kj} are realizations of the error random variables $\mathcal{E}_{kj}$. In theory, the weights w_{kj} are chosen to account for the variance of the $\mathcal{E}_{kj}$ following the assumptions of the statistical model. Thus, the statistical model has direct implications for the estimated best-fit parameter, given the data.

Two possible variance models were considered in Chapter 3. First, a constant variance (CV) statistical model was considered, in which case $Var(\mathcal{E}_{kj}) = \sigma_0^2$ and $w_{kj} = 1$ for all k and j. This results in the Ordinary Least Squares (OLS) framework (3.27). Second, a constant coefficient of variance (CCV) statistical model was considered, in which $Var(\mathcal{E}_{kj}) = \sigma_0^2 (I[\tilde{n}](t_j, z_k^j; \theta_0))^2$ and $w_{kj} = (I[\tilde{n}](t_j, z_k^j; \theta_0))^{-2}$. This results in the Generalized Least Squares (GLS) framework (3.28). In the absence of any a priori

knowledge regarding the correct form of the statistical model, the computationally simpler OLS model was used in Chapter 3 for the inverse problem.

In addition to the implications for confidence interval calculation discussed above, an accurate statistical model has implications for the weights w_{kj} in the inverse problem formulation (4.1). Additionally, the computation of the AIC values (Section 3.4.5) for modeling ranking and selection is premised upon modeling errors which are independent and normally distributed with constant variance. There is significant value, then, in ascertaining the properties of the error random variables and assessing the reliability of the assumptions made in the inverse problem procedure. In this chapter, the residuals r_{kj} which arise after the AIC selected best-fit model parameterization (A5B5dist; see Section 3.4.2) are examined in an effort to assess the reliability of the statistical model. As in [21], it is found that neither a CV nor a CCV assumption accurately describes the given data set. Moreover, it is found that the error random variables are not independent (given the current statistical model).

There are two potential causes for such observations. First, there is the possibility that the mathematical model does not accurately reflect the dynamics observed in the data. Second, it is possible that the statistical model (4.1) does not appropriately link the mathematical model with the data. In an effort to determine which is the case, several additional data sets have been collected so that variability in the experimental process could be assessed. It is shown (thankfully, perhaps) that the inaccuracy of the statistical model is the result of the failure of the statistical model to accurately incorporate sources of uncertainty which naturally arise in the experimental protocol of Chapter 1.

4.2 Analysis of Residuals

As discussed above, the residuals r_{kj} which result from fitting the model (here, the compartmental model from Chapter 3 with parameterization A5B5dist) to the data are realizations of the random variables $\mathcal{E}_{kj}$. As such, the reliability of the statistical error model can assessed by plotting the residuals r_{kj} and the modified residuals $r_{kj}/I[\tilde{n}](t_j, z_k^j; \theta_0)$ in terms of the model values $I[\tilde{n}](t_j, z_k^j; \theta_0)$. (In practice, of course, one must insert an estimate $\hat{\theta}_{OLS}$ or $\hat{\theta}_{GLS}$ in the place of the unknown θ_0 and use the approximate integral operator $I_A[\tilde{n}]$.) If a CV error model is sufficient to explain the noise in the data, then the residuals r_{kj} will be randomly distributed when plotted against the model values, while the variance of the modified residuals will decrease as the magnitude of the model increases. Alternatively, if a CCV model is sufficient to explain the noise in the data, then the modified residuals $r_{kj}/I_A[\tilde{n}](t_j, z_k^j; \theta_0)$ will be randomly distributed, while the original residuals will grow with the magnitude of the model. These observations are summarized in a hypothetical example in Figure 4.1. See [8, 22] for additional details and further examples. More details regarding the choice of statistical error and its effects on the inverse problem can be found in [8, 22, 37, 96].

Figure 4.2 shows the residuals and the modified residuals plotted in terms of the computed model values for the AIC-selected best-fit model parameterization A5B5dist. (Technically speaking, one should plot the modified residuals $r_{kj}/I_A[\tilde{n}](t_j, z_k^j; \theta_{GLS})$. However, the computation of the parameter estimate θ_{GLS} is quite expensive, and the model values $I_A[\tilde{n}](t_j, z_k^j; \theta_{GLS})$ would change minimally.) When the residuals are plotted in terms of the value of the observed model solution there is a clear increase in the variance of the residuals as the size of the model increases, providing an indication that the assumption

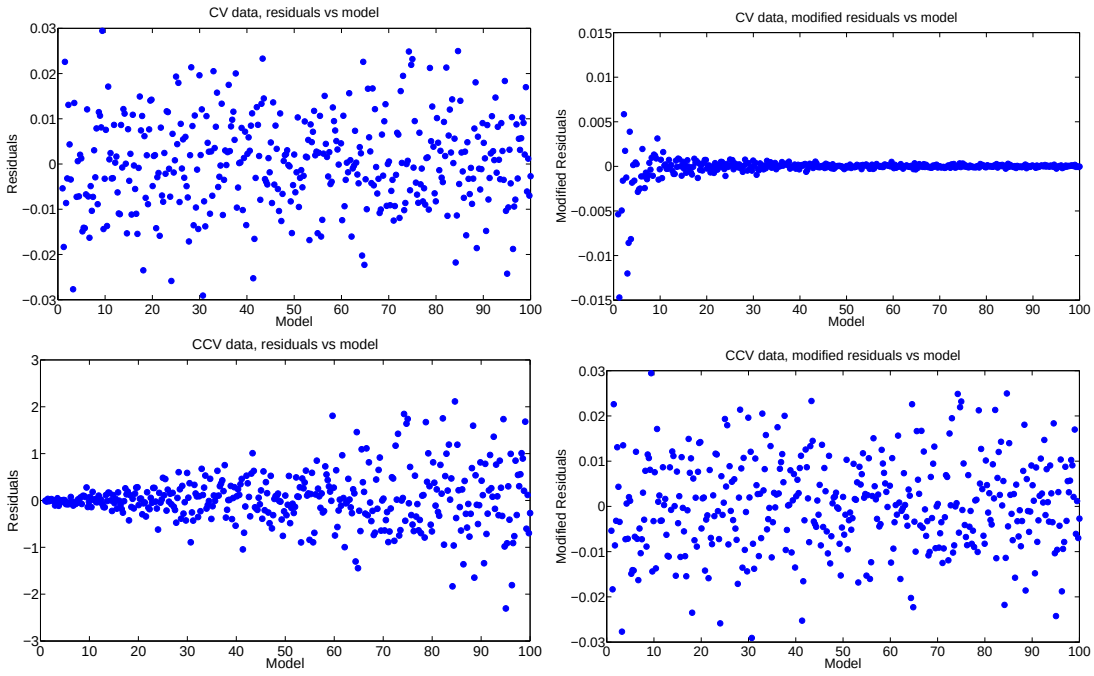

Figure 4.1: Top: Hypothetical residuals (left) and modified residuals (right) for constant variance (CV) data when plotted in terms of the model value. Bottom: Hypothetical residuals (left) and modified residuals (right) for constant coefficient of variation (CCV) data when plotted in terms of the model value. When the correct statistical model is used (top-left and bottom-right), the residuals (or modified residuals) appear randomly distributed. The fan-like structures in the top-right and bottom-left panels are characteristic of such residual plots when the statistical model for the measurement errors has been misspecified.

of CV errors may be incorrect. However, the residuals also lack the fan-like structure typical of CCV errors (Figure 4.1). When the modified residuals are plotted in terms of the magnitude of the observed model solution, the pattern is distinctly nonrandom. Thus, it appears that the true statistical model for the errors may lie somewhere between the CV model (OLS estimation) and the CCV model (GLS estimation), perhaps slightly closer to the CV model.

The assumption that the error random variables at each of the data points are independent may also be problematic. For instance, when the residual plots are separated in terms of measurement times (Figure 4.3) additional structure is noticeable in the residual plots when compared to Figure 4.2. The independence of the error random variables can be investigated with a scatterplot of the residuals r_{kj} (which are considered as realizations of those random variables) in terms of the previous residual $r_{k(j-1)}$. If the error random variables were truly independent, then such a scatterplot would have no discernable structure. However, we see in Figure 4.4 that this is not the case, as there is a clear positive correlation between the sets of residuals.

There are two possibilities which may explain the positive correlation between the two sets of residuals. First, it is possible that neighboring data points are not independent. Because the data used to

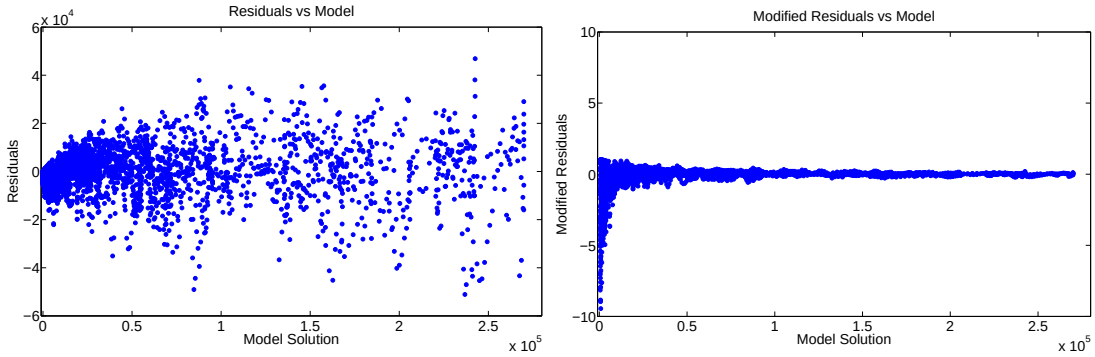

Figure 4.2: Residuals (left) and modified residuals (right) for the OLS best-fit model solution. Because neither graphic exhibits constant variance, the assumptions of CV error and CCV error must both be wrong. While the misspecification of the error term does not invalidate the ability of the compartmental model to fit the data, we cannot determine the statistical properties (e.g., confidence intervals) of the parameter estimates.

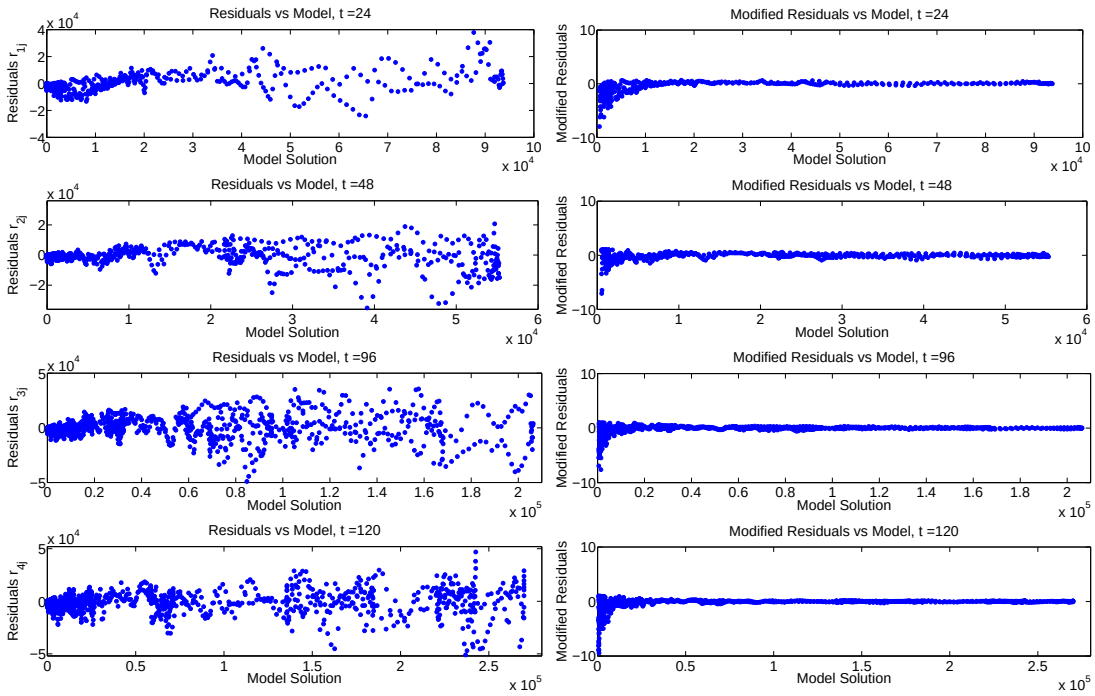

Figure 4.3: Residuals (left) and modified residuals (right) for the OLS best-fit model solution, shown separately for each measurement time. There is some additional structure which is evident in these residual plots which is not evident when the residuals for all measurement times are shown together (Figure 4.2). This has implications for the statistical model of the data.

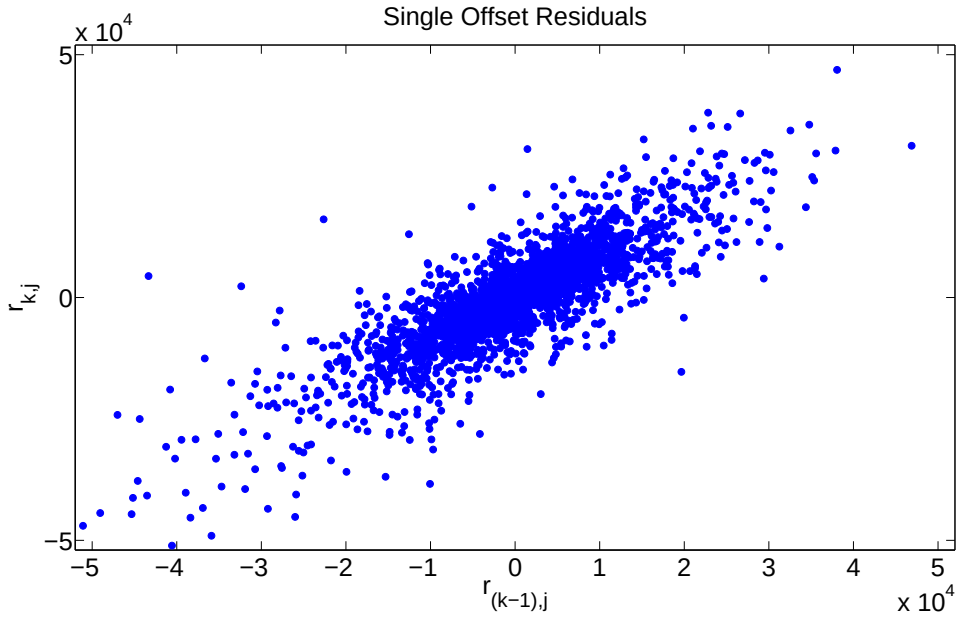

Figure 4.4: The assumption of independent errors for each observation can be checked by plotting the residuals r_{kj} against the offset residuals $r_{k(j-1)}$. If the errors were independent, there would be no discernable structure in such a graphic. However, we observe a clear positive correlation.

calibrate the model is histogram data, it is possible that the number of cells counted into adjacent bins (and hence, the error terms) might be linked by the location of the boundary separating the adjacent bins. In general, this might be demonstrated by investigating how the noise in the data changes as the bins used to generate the histogram data changes. Unfortunately, the data set used in this report was received with the bins already fixed. Still, these effects deserve careful consideration and must be addressed in future work. Some research has also indicated that cells which descend from a common precursor may share certain traits and/or behaviors [58, 106]. It is unclear how such correlation might impact either the error terms or the mathematical model itself, though this is considered briefly in Chapter 5.

A second possibility is that the model, though close to the data (see, e.g., Figure 3.9) does not satisfy the tacit assumption $E[N_k^j] = I[\tilde{n}](t_j, z_k^j; \theta_0)$ (or, equivalently, $E[\mathcal{E}_{kj}] = 0$). Given the convergence properties of the numerical solution discussed in Chapter 3, it seems unlikely that the failure of this assumption could be caused by any computational errors or approximations. A more likely explanation is the discrepancy between the assumptions regarding the collected data and the actual experimental reality. As discussed in Section 1.2, the 5 samples of data collected (4 used for fitting the model plus one for the initial condition) are actually 5 separate samples taken from the same donor. While each sample receives an identical treatment, the assumption that all five samples are identical for all time may be inaccurate. Moreover, only a fraction of each sample is measured, and a scaling factor (which may also be subject to error) is used to adjust the resulting cell counts. Meanwhile, the model is derived under the assumption that each histogram represents a complete census of the cells in the population,

and that the same population (i.e., cells arising from the same set of precursors) is measured each time. In order to correct for such a discrepancy between the assumptions of the mathematical model and the experimental reality, a more rigorous, detailed observation operator may be needed which accounts for the experimental sampling method, with its attendant sources of error. This may also help to resolve the slight negative bias of the residuals observed in Figure 4.2.

Given these two possibilities, it is unavoidable that the natural variability in the experimental protocol must be assessed. Rather than continue to use a single data set to investigate the two possibilities above (and risk 'overfitting' the data [31]), several additional data sets have been collected for these purposes.

4.3 Data Variability

As discussed above, the work presented thus far on modeling CFSE/flow cytometry data has been based upon several assumptions. It is tacitly assumed that at each measurement time, the histogram data (to which the mathematical model is fit) is representative of a total census of the cell population. That is, it is assumed that the sample of cells measured from each well is representative of the population of cells in the well, and that the sampled fraction of cells can be accurately estimated using counted beads. It is also assumed that the same population of cells (i.e., cells arising from the same set of precursors) is used at each measurement time. While experimental protocol necessitates the use of cells from separate well plates at each measurement time, it was hoped that the populations of cells in the individual well plates were sufficiently alike so that this assumption would be mathematically reasonable.

However, it has already been acknowledged in Chapter 1 that certain data sets appear to violate these assumptions. For example, the data collected at $t = 72$ hours in the present data set could not be used because the increase in the number of cells from the previous measurement was physically impossible (given the assumptions on the data). It should be emphasized that this failure is not unique to the current mathematical efforts as all analyses of CFSE data (e.g., the precursor-cohort method, the cyton model, etc.–see Chapter 1 for extensive references) are at least tacitly premised upon the assumption that CFSE histogram data represents a census of a single cell population. It should also be emphasized that the violation of the assumptions on the data is *not the result of an experimental failure*. Rather, there is a natural experimental variability to be expected *when a sample of the population is measured as a means of approximating the behavior of the total population*. Previous authors have noted the importance of considering experimental variability in CFSE data sets [109]. Because one significant goal of the mathematical modeling of flow cytometry data is the use of parameter estimates to compare various experimental and biological conditions, there is also a need to examine how the data varies between two nominally healthy individuals, as well as how the data varies from sample to sample within the same donor. That is, the intra- and inter-individual variability in CFSE data sets must be considered in addition to the variability in the experimental procedure.

In order to address these issues, multiple data sets have been collected for the purpose of quantifying variability in flow cytometry data. In order to assess inter-individual variability between two healthy donors, cells were collected from two donors (Donor 1 and Donor 2). A fraction of cells from each donor was set aside (in order to measure the AutoFI properties of the cells) while the remaining cells were labeled with CFSE. For both CFSE^+ and CFSE^- cells, half of the cells were stimulated with PHA

while the other half of the cells were not stimulated. Cells were placed into individual well plates at a density of approximately 0.5×10^6 cells per well. In order to assess experimental variability, cells labeled with CFSE were measured in triplicate on days 1, 3, and 5. Thus, for each donor, there were 3 wells for $CFSE^-PHA^-$ cells, 3 wells for $CFSE^-PHA^+$ cells, 9 wells for $CFSE^+PHA^-$ cells, and 9 wells for $CFSE^+PHA^+$ cells. At each measurement time, flow cytometry data was collected on

- unstimulated AutoFI ($CFSE^-PHA^-$; one well per day),
- stimulated AutoFI ($CFSE^-PHA^+$; one well per day),
- unstimulated, labeled cells ($CFSE^+PHA^-$, three wells per day),
- stimulated, labeled cells ($CFSE^+PHA^+$, three wells per day).

Gating procedures were used to isolate the CD4+ cells, which were stored for the analysis presented here. In order to address intra-individual variability, this entire experiment was repeated 17 days later (henceforth, we refer to the two experiments as Trial 1 and Trial 2) with the same two donors. It is hoped that a qualitative analysis of these data sets can be used as a basis for modifying the statistical model which relates the data to the mathematical model. Once accomplished, the appropriate incorporation of experimental uncertainty into the statistical model will permit a more meaningful (and more accurate) estimation of parameters without the need for assumptions (on the experimental data) which are of limited validity. The analysis of AutoFI is postponed until the next chapter.

4.3.1 Experimental and Donor Variability

As has been previously noted, the original data set used in this research exhibited the 'precursor cohort problem', in which the total number of precursors (the cells originally in the population which give rise to all other cells after division) or the total quantity of CFSE FI (see Figure 1.1) increases from one measurement time to the next. If it were safe to assume that each measurement can be treated as if it were a census of the same population of cells, then such an increase in the total number of precursors would not be possible, as it violates the conservation law upon which the mathematical model is premised. However, it is known that only a fraction of cells in a given well are counted by the flow cytometer, and that a different well is used for each measurement time. Data collected from such samples does not pose a theoretical problem in itself, provided a known fraction of the total population is sampled, and that the sampled cells are representative of the total population. The fraction of cells measured out of each well is estimated by placing a known number of beads (in the data sets for this document, 51175) into each well. These beads have well-known fluorescence properties and can be counted in the flow cytometry data. The estimated total number of cells in a well is determined by scaling the number of measured cells by the ratio of total beads to counted beads. More than likely, the number of measured cells or the number of counted beads (or both) is subject to some measurement error. It is also likely that the number of cells placed into distinct wells may vary as well. Thus, it seems that the precursor cohort problem might be explained by natural variability in the experimental procedure.

In order to investigate these sources of error, Tables 4.1 and 4.2 show the numbers of measured cells, counted beads, and estimated total number of cells for the CFSE labeled unstimulated cells from Donors

Table 4.1: Summary of event counts for Donor 1 without stimulation. In two separate experiments occurring 17 days apart, triplicate samples of cells were measured on days 1, 3, and 5. The number of cell events measured by the cytometer, the number of bead events measured by the cytometer, and the resulting estimated total cell population are shown for each measurement. Data sets in which total cell numbers increase between two subsequent measurement times are highlighted in yellow. For instance, in Trial 1, the total number of cells in the first sample increases from 301159 to 375386 between days 3 and 5.

	Trial 1			Trial 2		
	Cell Events	Bead Events	Total Cells	Cell Events	Bead Events	Total Cells
Day 1	19906	2116	481422	20136	2525	408103
	19485	2330	427959	20307	2869	362221
	19844	2360	430304	20501	2305	455158
Day 3	21062	3579	301159	21105	2322	465137
	21372	2571	425403	21390	2364	463043
	20860	2715	393190	21353	2429	449872
Day 5	22006	3000	375386	22515	2986	385869
	21667	3069	361293	22848	3201	365275
	21520	3109	354225	23827	4079	298933

Table 4.2: Summary of event counts for Donor 2 without stimulation. As in Table 4.1, in two separate experiments, triplicate samples of cells were measured on days 1, 3, and 5. The number of cell events measured by the cytometer, the number of bead events measured by the cytometer, and the resulting estimated total cell population are shown for each measurement. Data sets in which total cell numbers increase between two subsequent measurement times are highlighted in yellow.

	Trial 1			Trial 2		
	Cell Events	Bead Events	Total Cells	Cell Events	Bead Events	Total Cells
Day 1	22498	2015	571382	18451	2392	394745
	22992	1913	615063	17892	2024	452383
	22756	1990	585195	17787	2474	367926
Day 3	24040	2157	570351	20120	2367	434998
	24006	2278	539292	20338	2458	423433
	24088	2352	524109	20677	2490	424958
Day 5	25050	3367	380735	23398	2659	450317
	24921	3500	364381	23199	3159	375818
	25186	3048	422865	22797	2715	429700

Table 4.3: Summary of the estimated total cell numbers (hence total precursor numbers) for the CFSE labeled unstimulated cells. Of the 12 data sets collected, 5 (highlighted in yellow) exhibit an increase in the total number of precursors from one day to the next.

Data Set	Day 1	Day 3	Day 5
Donor 1 Trial 1 Sample 1	481422	301159	375386
Donor 1 Trial 1 Sample 2	427959	425403	361293
Donor 1 Trial 1 Sample 3	430304	393190	354225
Donor 1 Trial 2 Sample 1	408103	465137	385869
Donor 1 Trial 2 Sample 2	362221	463043	365275
Donor 1 Trial 2 Sample 3	455158	449872	298933
Donor 2 Trial 1 Sample 1	571382	570351	380735
Donor 2 Trial 1 Sample 2	615063	539292	364381
Donor 2 Trial 1 Sample 3	585195	524109	422865
Donor 2 Trial 2 Sample 1	394745	434998	450317
Donor 2 Trial 2 Sample 2	452383	423433	375818
Donor 2 Trial 2 Sample 3	367926	424958	429700

1 and 2, respectively (note that each measurement was made in triplicate). Because these cells are not stimulated, the cells will not divide and the total number of cells in the population is equal to the total number of precursors in the population. Thus, one can examine the precursor cohort problem directly without any need to fit the data or use a deconvolution technique in order to determine the number of cells in each generation. If one were to make the common assumption that the cells in each well are identical, then the estimated total number of cells cannot increase from one day to the next (though the number may decrease as a result of cell death). For Donor 1, one of the triplicate data sets from Trial 1 features estimated total cell numbers which increase. In Trial 2, two of the three triplicate data sets have this precursor cohort problem. The problematic samples are highlighted in yellow.

Thus, between two donors and two trials, a total of 12 time-series data sets were obtained for unstimulated cells. In 5 of those 12 data sets, the precursor cohort problem appears. These twelve data sets, given in detail in Tables 4.1 - 4.2, are summarized in Table 4.3. It should be emphasized that because these cells were not stimulated to divide, there is no deconvolution or fitting of the data involved in determining the number of precursors; the increase in precursors from one day to the next must be experimental and not mathematical. It should also be emphasized that this is not an experimental error and almost certainly cannot be avoided. There is a natural variability in the experimental data as a result of the necessary steps required to analyze a population of cells based upon measurements obtained from only a sample of those cells. It is of interest, however, to understand the potential causes of such variability so that they can be addressed mathematically or statistically.

There are three potential sources of variability underlying the precursor cohort problem. First, it is possible that the number of counted beads is subject to error. This may be the case if some particles and/or cells in the culture have fluorescence properties similar to that of the beads. Then, when the gates are set to isolate and count the measured beads, the resulting number will be biased upward. The magnitude of such bias will depend upon the number of such cells/particles. Alternatively, if the spectral properties of the beads are not sufficiently similar (or are subject to significant measurement error) the

Table 4.4: Summary of event counts for Donor 1 for a population stimulated with PHA. In two separate experiments occurring 17 days apart, triplicate samples of cells were measured on days 1, 3, and 5. The number of cell events measured by the cytometer, the number of bead events measured by the cytometer, and the resulting estimated total cell population are shown for each measurement. Unlike in Table 4.1, these cells are dividing and thus the observed increases in the total number of cells is expected.

	Trial 1			Trial 2		
	Cell Events	Bead Events	Total Cells	Cell Events	Bead Events	Total Cells
Day 1	20799	3984	267166	20175	4953	208451
	20487	3855	271964	19872	4104	247795
	20914	3731	286860	20322	4027	258251
Day 3	18465	3155	299508	17798	2602	350043
	18546	3340	284159	17394	2537	350862
	19547	4396	227552	17315	2855	310366
Day 5	13722	1003	700123	14255	1051	694100
	13539	1079	642130	15436	905	872859
	13654	1204	580352	15665	1056	759144

Table 4.5: Summary of event counts for Donor 2 for a population stimulated with PHA. As in Table 4.4, in two separate experiments occurring 17 days apart, triplicate samples of cells were measured on days 1, 3, and 5. The number of cell events measured by the cytometer, the number of bead events measured by the cytometer, and the resulting estimated total cell population are shown for each measurement.

	Trial 1			Trial 2		
	Cell Events	Bead Events	Total Cells	Cell Events	Bead Events	Total Cells
Day 1	24923	3617	352622	19438	3813	260881
	24885	2778	458420	19588	3790	264490
	24793	3321	382048	19847	3528	287888
Day 3	19698	2132	472817	14101	3434	210139
	19577	2400	417439	13291	2964	229476
	19619	2648	379155	14077	2576	279655
Day 5	15375	1051	748635	10961	1213	462431
	14908	1051	725896	11014	1193	472457
	14803	938	807616	10850	1250	444199

beads will appear spread out in a scatterplot. In that case, the number of beads gated and counted will be biased downward because some beads which were present in the data were excluded by the gates.

A second possibility is that the number of cell events counted in the cytometry data may be subject to error. Because the measured culture contains a wide variety of cells, various properties of the cells of interest (here, CD4+ cells) must be used (e.g., size, granularity, CD4 expression) in order to isolate those cells. However, each gate is a potential source of error. The magnitude and direction of those errors will depend upon variability in the fluorescence properties of the cells of interest as well as on the proximity of other cells (in a scatterplot) to the regions containing the cells of interest. Similarly, because cell duplets must be removed from the data, this may be a source of systematic error in the computed cell numbers. For instance, if there are a large number of CD4+ cells clumped together at the beginning of the experiment, these cells will be removed from the data. Then, if cells tend to unclump during the course of the experiment, fewer cells will be removed, leading to an apparent increase in the number of cells in the population. The third possibility is variability in the number of cells initially placed into the individual wells. If different wells were to receive different numbers of cells at the beginning of the experiment, one would naturally expect some variability in the number of precursors computed from the data!

Given these three potential sources of experimental variability, we return to the data in Tables 4.1 and 4.2. Notice that the number of cell events, though different for the two donors and the two trials, is very similar for each set of triplicate measurements. (The number of measured cell events for Donor 1 Trial 2 Day 5 has a relative difference of less than 6% among the triplicate samples, and all remaining measurements feature differences of less than 3%.) This would seem to indicate that the variability in the numbers of counted cells is small. The number of counted beads, however, can vary much more significantly (as much as 33% relative difference for Donor 1 Trial 1 Day 3). Thus it is the variation in the numbers of counted beads which seems to be more responsible for the observed precursor cohort problems. Because we do not have any means (using only the current data sets) to analyze fluctuations in the numbers of cells initially placed into each well, these conclusions must be taken with some skepticism. For instance, we must leave open the possibility that the numbers of cell events and bead events are quite accurate, so that it is the actual true number of cells in each well which is variable. In order to determine which of the three possibilities is the most accurate account of the experimental variability, more information regarding the data preprocessing steps (the gating procedures) is needed, and this analysis is ongoing.

Of course, data collected for unstimulated cells is quite static. Because our primary focus is on modeling dynamic populations which are activated and dividing, we would like to see if the observations noted above hold true for PHA-stimulated cells as well. In Tables 4.4 and 4.5, the counted cells, beads, and computed total cells are shown for Donors 1 and 2 for cells stimulated with PHA. Note that, for PHA-stimulated cells, proliferation will occur so that increases in total cell count are permissible. Just as for unstimulated cells, the data in Tables 4.4 and 4.5 exhibit consistent numbers of cell events in the triplicate data sets, with more variability in the numbers of counted beads.

Regardless of the exact causes, it seems safe to conclude that natural experimental variability is responsible for observed fluctuations in the total numbers of cells estimated to be in the measured population. Beyond this variability, there is great interest in quantifying intra-individual variability–

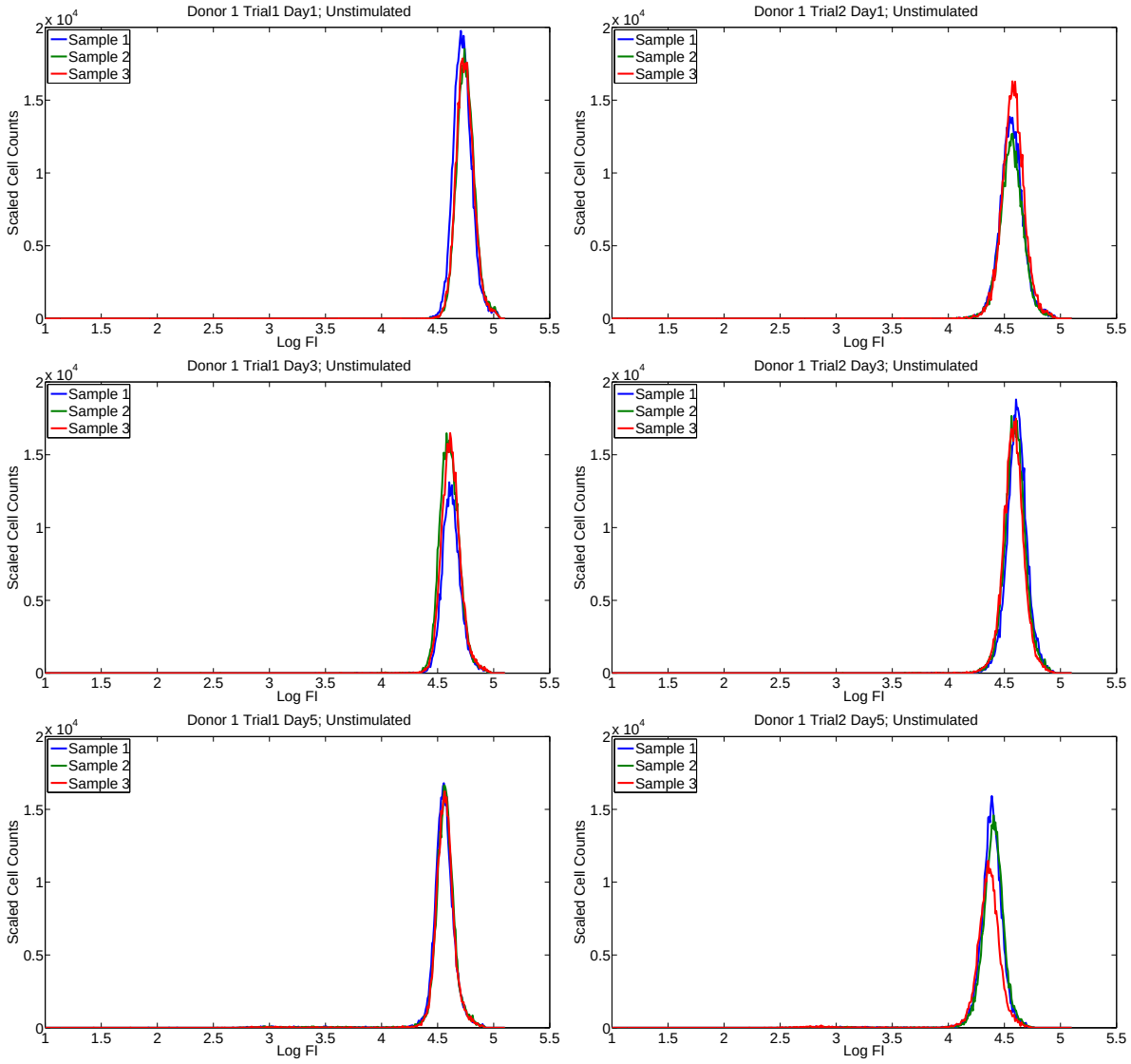

Figure 4.5: Data from Donor 1, cells unstimulated. Left: Trial 1 data, collected in triplicate, on days 1, 3, and 5. Right: Trial 2 data, taken 17 days after the conclusion of the previous experiment. The cell counts for these data sets can be found in Table 4.1.

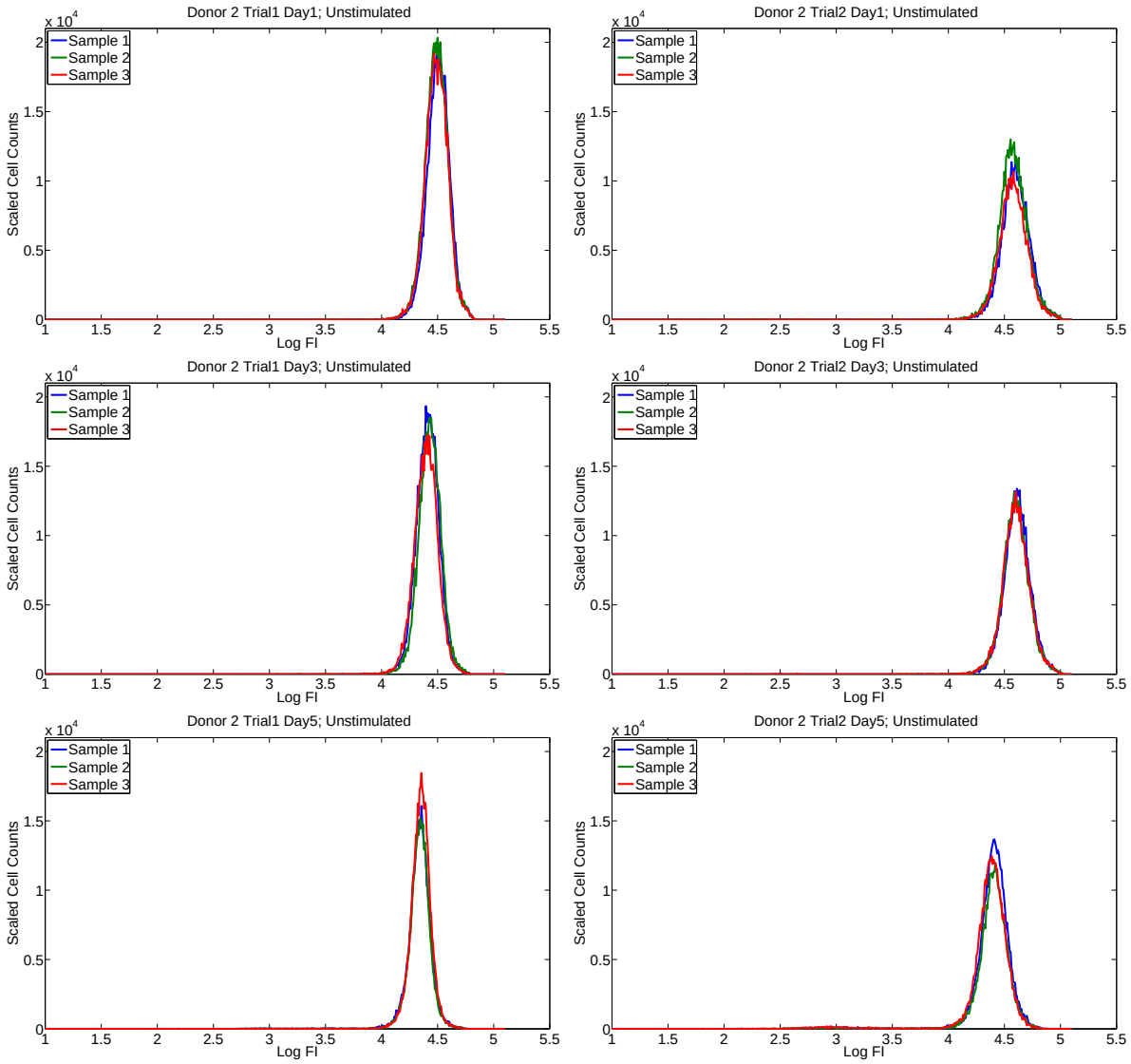

Figure 4.6: Data from Donor 2, cells unstimulated. Left: Trial 1 data, collected in triplicate, on days 1, 3, and 5. Right: Trial 2 data, taken 17 days after the conclusion of the previous experiment. The cell counts for these data sets can be found in Table 4.2. While the total number of cells is occasionally variable, the overall shape of the histogram data is consistent across the triplicate samples and from one trial to the next.

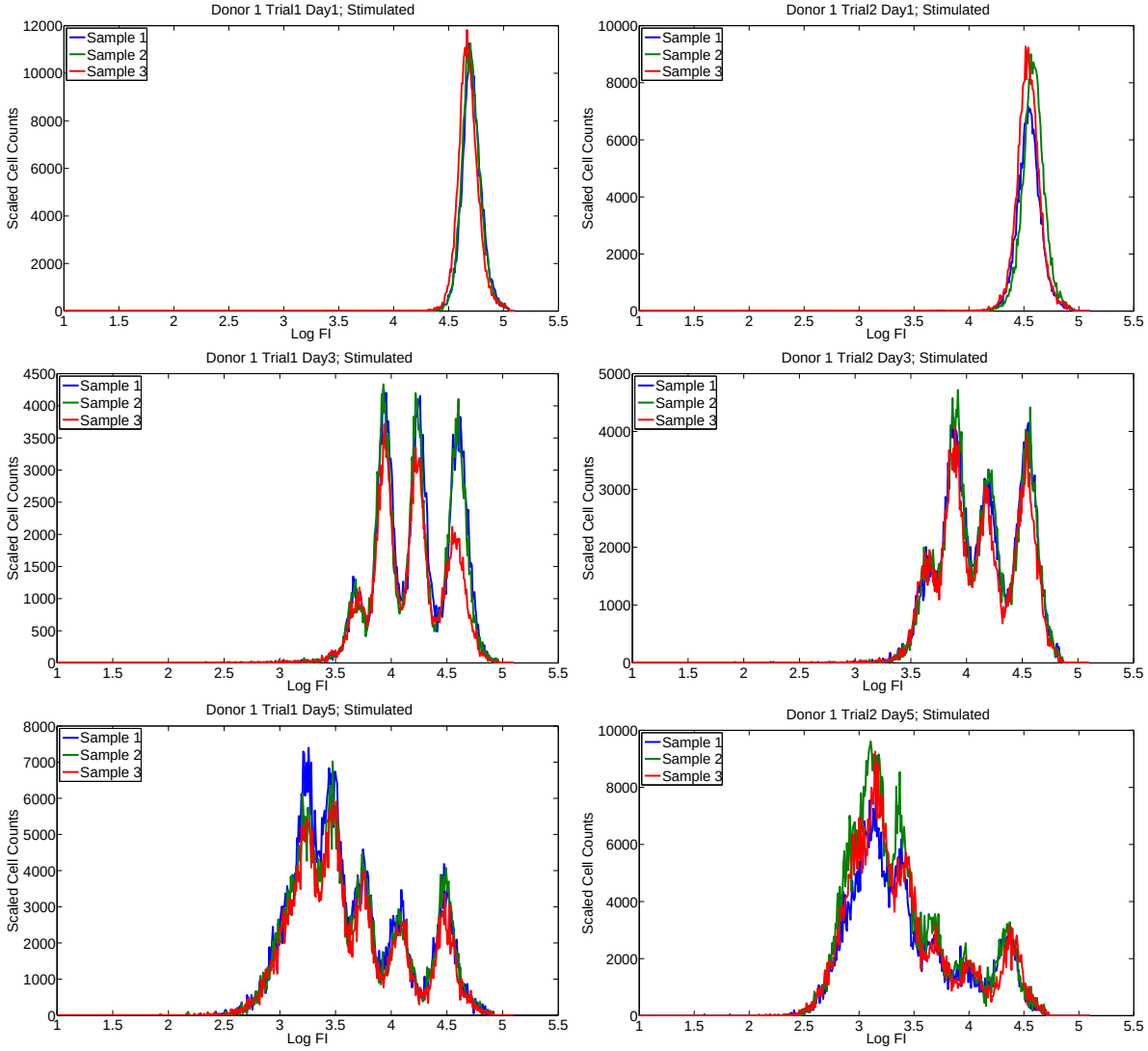

Figure 4.7: Data from Donor 1, cells stimulated with PHA. Left: Trial 1 data, collected in triplicate, on days 1, 3, and 5. Right: Trial 2 data, taken 17 days after the conclusion of the previous experiment. The cell counts for these data sets can be found in Table 4.4. With the exception of the data from Trial 1 Day 3, the overall shape of the histogram data is consistent across the triplicate samples and from one trial to the next. The cause of the unexpectedly small number of undivided cells in the histogram data for Trial 1 Day 3 Sample 3 is unknown–it may be biological or experimental.

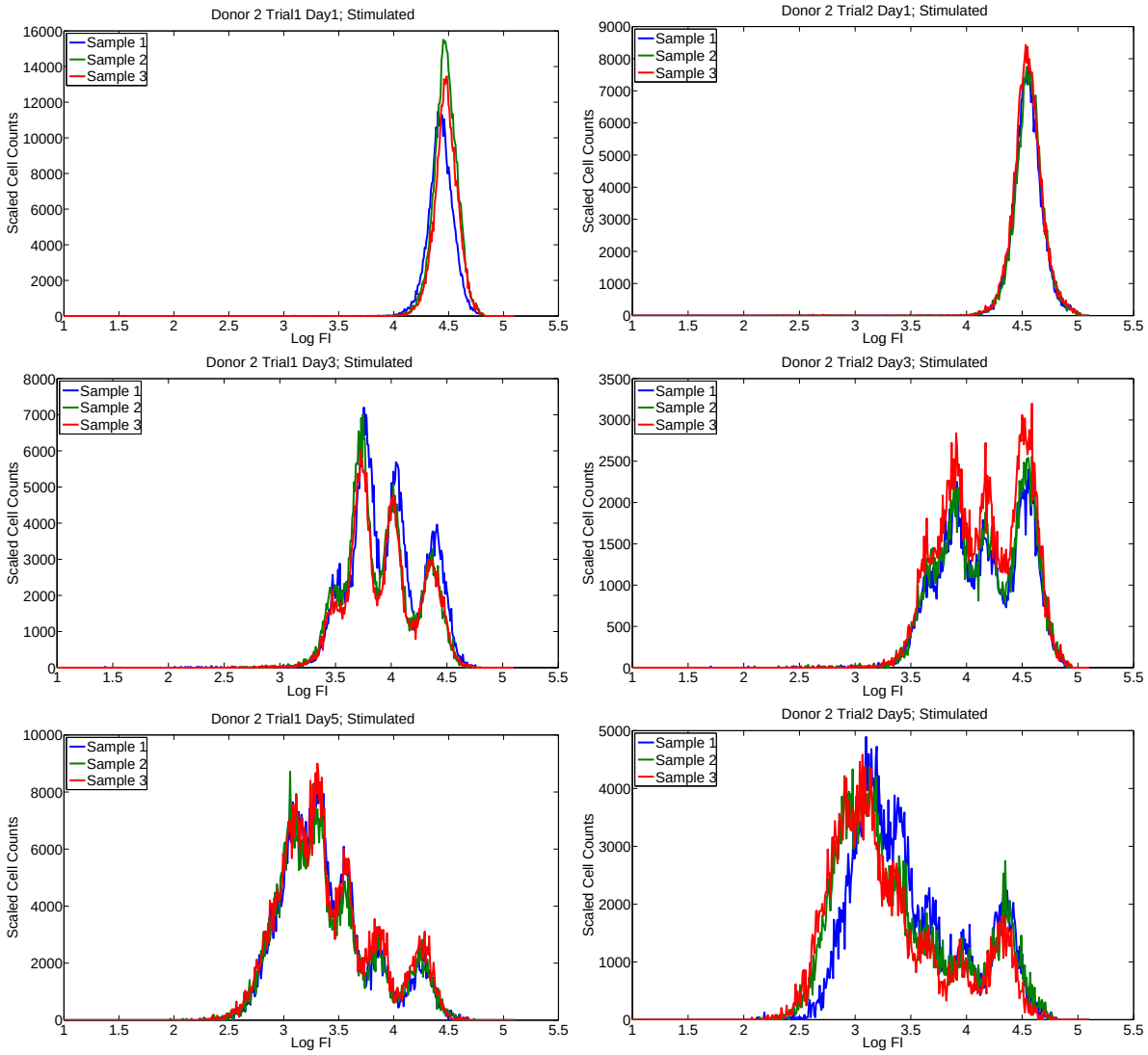

Figure 4.8: Data from Donor 2, cells stimulated with PHA. Left: Trial 1 data, collected in triplicate, on days 1, 3, and 5. Right: Trial 2 data, taken 17 days after the conclusion of the previous experiment. The cell counts for these data sets can be found in Table 4.5. With the exception of the data from Trial 2 Day 5, the overall shape of the histogram data is consistent across the triplicate samples and from one trial to the next. The cause of the additional generation of cells in the histogram data for Trial 2 Day 5 Sample 3 is unknown–it may be biological or experimental.

that is, how much variability is there in the *behavior* of cells, even when those cells were taken from the same donor. In order to do so, we look at the CFSE histogram data formed from the cells measured for each donor. Because each measurement is made in triplicate, and because two separate experiments were conducted 17 days apart, we can analyze differences in cell behavior in order to understand variability within a single individual. We make the reasonable assumption that cells which behave in a sufficiently similar manner will result in CFSE histograms which are similar in overall shape. It should be noted that the CFSE histograms are computed from flow cytometry data which we have already acknowledged to be subject to experimental variability. However, as this variability affects the *numbers* of cells estimated to be in the population, and not the actual *behavior* of those cells, it follows that histograms which are similar in shape (but possibly not in size) are indicative of similar populations of cells.

We begin with the histogram data for the unstimulated cells (summarized in Tables 4.1 and 4.2). Figures 4.5 and 4.6 show the histogram data for the unstimulated cells from Donors 1 and 2, respectively. We see in these figures that, aside from the sample-to-sample variability in the total number of cells (which we know from the previous subsection to be problematic), the data are quite similar from one trial to the next, and from sample to sample within the same trial and donor. Of course, these cells have not been stimulated and it is of little surprise that such static populations result in similar histograms.

Thus, we turn our attention to the PHA-stimulated cells which were summarized in Tables 4.4 and 4.5. Figures 4.7 and 4.8 depict the data sets for Donors 1 and 2 for these cells. When ignoring differences in total computed cell numbers we see that even these dynamic data sets are generally consistent across triplicate samples for a donor. For instance, the relative sizes of each of the peaks (hence the approximate percentage of cells in each generation) is consistent across the various samples measured at each day. The only exceptions are the data sets for Donor 1 Trial 1 Day 3 (Figure 4.7) and Donor 2 Trial 2 Day 5 (Figure 4.8), in which one of the triplicate measurements is noticeably different from the other two. For Donor 1 Trial 1 Day 3, there are far fewer undivided cells in the Sample 3 data than would be expected given the other two samples. For Donor 2 Trial 2 Day 5, one additional generation of cells is clearly visible in the data from Sample 3 when compared to the other two samples. It is unclear what may be the cause of these two instances of intra-individual variability. It is possible that subtle experimental differences in the populations of cells used for the triplicate samples (e.g., variations in nutrient concentration, strength of stimulation, exact time of measurement, etc.) can account for these variations. It is also possible that elements of the measurement process (e.g., a subset of cells gated in or out of the population) could be responsible.

There are some noticeable differences in Figures 4.7 and 4.8 between the measured populations of cells in Trial 1 compared to those in Trial 2 (compare the left and right columns of graphics). While there are generally the same numbers of generations of cells visible on a given day, the distribution of cells among different division numbers is different between the two trials. These differences are probably explained by subtle variations in the measurement process between the two sets of data. Data for Trial 1 was collected 22.5, 69.5, and 115.5 hours after stimulation. Data for Trial 2 was collected 20, 69, and 118 hours after stimulation. These slight differences in timing seem sufficient to explain the subtle differences in the distribution of cells among generations between Trial 1 and Trial 2.

While there does appear to be some intra-individual variability in CFSE histogram data, this variability appears to be less frequent and smaller in magnitude than the experimental variability associated

with the total number of cells in the population. Given the similarity of the triplicate samples obtained for each donor, as well as the similarity of the data between Trial 1 and Trial 2 for each donor, it seems reasonable to conclude that, for a given mathematical model (e.g., the compartmental model in Chapter 3) the parameters of that model can be consistently estimated for a given healthy individual. Future work may be directed toward fitting the mathematical model to each triplicate data set (for each donor and trial) and examining differences in the estimated parameters. As noted at the beginning of this document, however, we must first identify an accurate statistical model which links the mathematical model to the data (see Section 4.5). Second, we must determine the degree to which parameters in the mathematical model can be uniquely estimated, which will almost certainly depend upon the times at which measurements are taken.

Given the results so far established concerning levels of experimental and intra-individual variability, we next consider how CFSE histogram data differs between the two healthy donors. *One of the ultimate goals of modeling CFSE histogram data is to use mathematical parameters estimated from a data set to distinguish between diseased and healthy cells.* As such, we must establish how much variation can be expected between data sets for two different but nominally healthy donors. As noted above, differences in the actual mathematical parameters (when a model is fit to multiple data sets) will depend upon both an accurate statistical model as well as issues of uniqueness in the inverse problem solution. These discussions are postponed for the moment. We can, however, examine the overall shapes of the histograms obtained from measurements of cells from the two donors. If the cells from the two donors behave in sufficiently similar manners, then the histograms should have similar shapes. In Figure 4.9, histogram data for unstimulated cells from the two donors on day 1 are shown in comparison to one another. (It follows from Figures 4.5 and 4.6 that the histograms for days 3 and 5 are nearly identical to those for day 1, so we do not show these.) As above, there is some experimental variability in the total number of cells estimated to be in each population. But because the current mathematical model (from either Chapter 2 or Chapter 3) is linear, variations in total cell numbers between donors are irrelevant. (Differences in total cell numbers do still matter from one day to the next within the same donor.) It is plainly observed that cells from Donor 1 were more brightly labeled with CFSE than cells from Donor 2 during Trial 1. This difference is also inconsequential as the initial distribution of CFSE is set independently for each data set when the compartmental model is calibrated to data. The shape of the CFSE distributions, as well as the rate at which CFSE FI decreases, is approximately the same for both donors once differences in total cell numbers and initial labeling intensity are accounted for.

As before, there is little information contained in the data sets for unstimulated cells. We are primarily interested in quantifying differences in dynamic behavior between the two donors. Thus, the CFSE histograms for the PHA-stimulated cells from the two donors are shown in comparison for each day in Figures 4.10 - 4.12. Again, we observe the expected sample-to-sample and donor-to-donor variation in the total number of cells in the population. As in the unstimulated case (Figure 4.9), cells for Donor 1 obtained a brighter initial staining than cells for Donor 2 during Trial 1. However, after accounting for these differences, the overall shape of the data sets (e.g., the proportion of cells in each generation) is relatively consistent for the two donors. This provides some hope that it may be possible to establish a narrow set of parameters (in the mathematical model) which can reliably fit data from healthy donors. In addition to the immediate advantage of reducing the computational time associated with fitting a

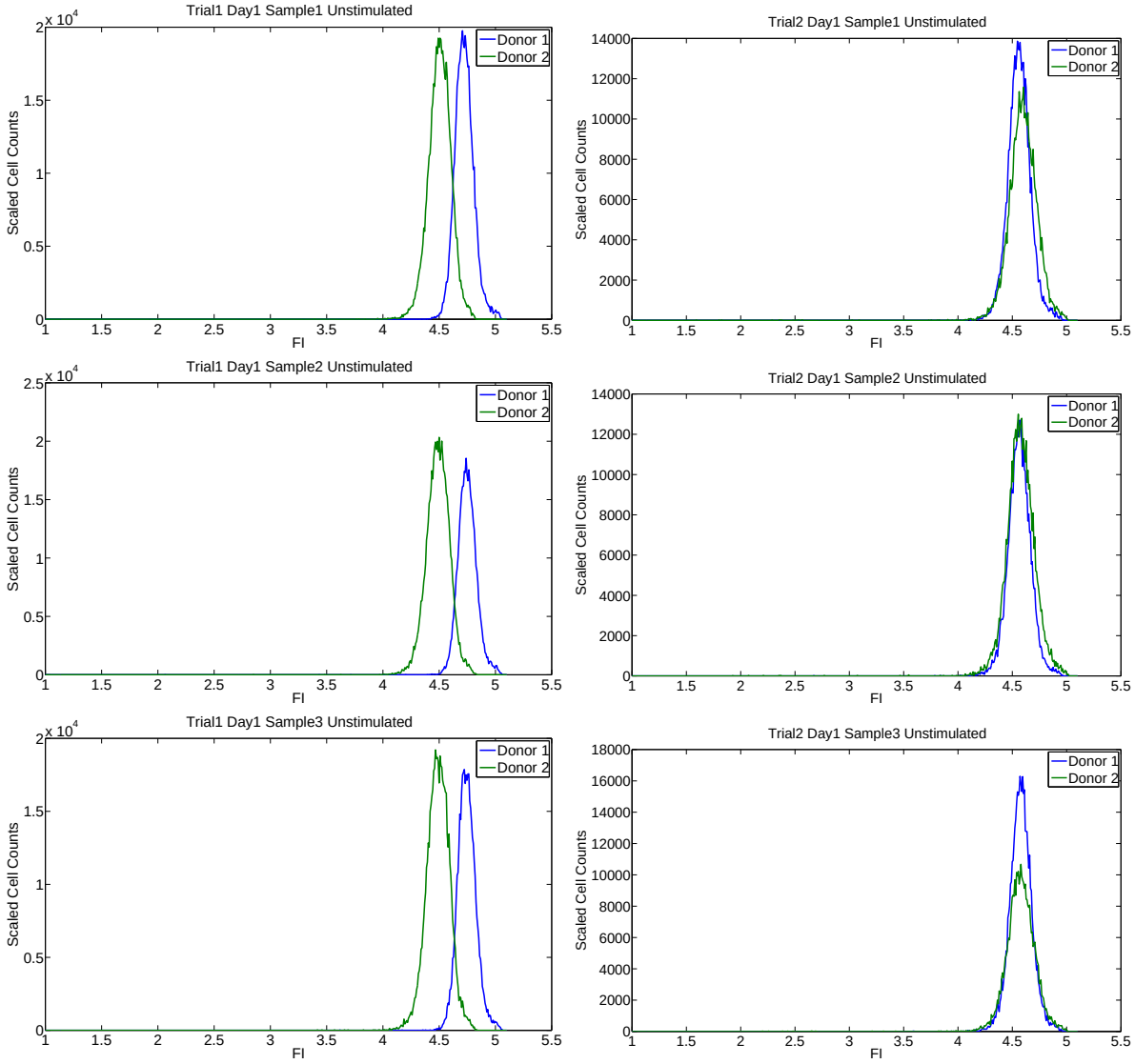

Figure 4.9: Inter-individual variability among unstimulated cells at day 1. Left: Trial 1. Right: Trial 2. Data for days 3 and 5 is sufficiently similar for unstimulated cells and thus is not shown. Cells from Donor 1 received a brighter initial labeling of CFSE than cells from Donor 2 during Trial 1. This difference will not affect the parameters associated with the dynamic behavior (e.g., proliferation and death) of the cells.

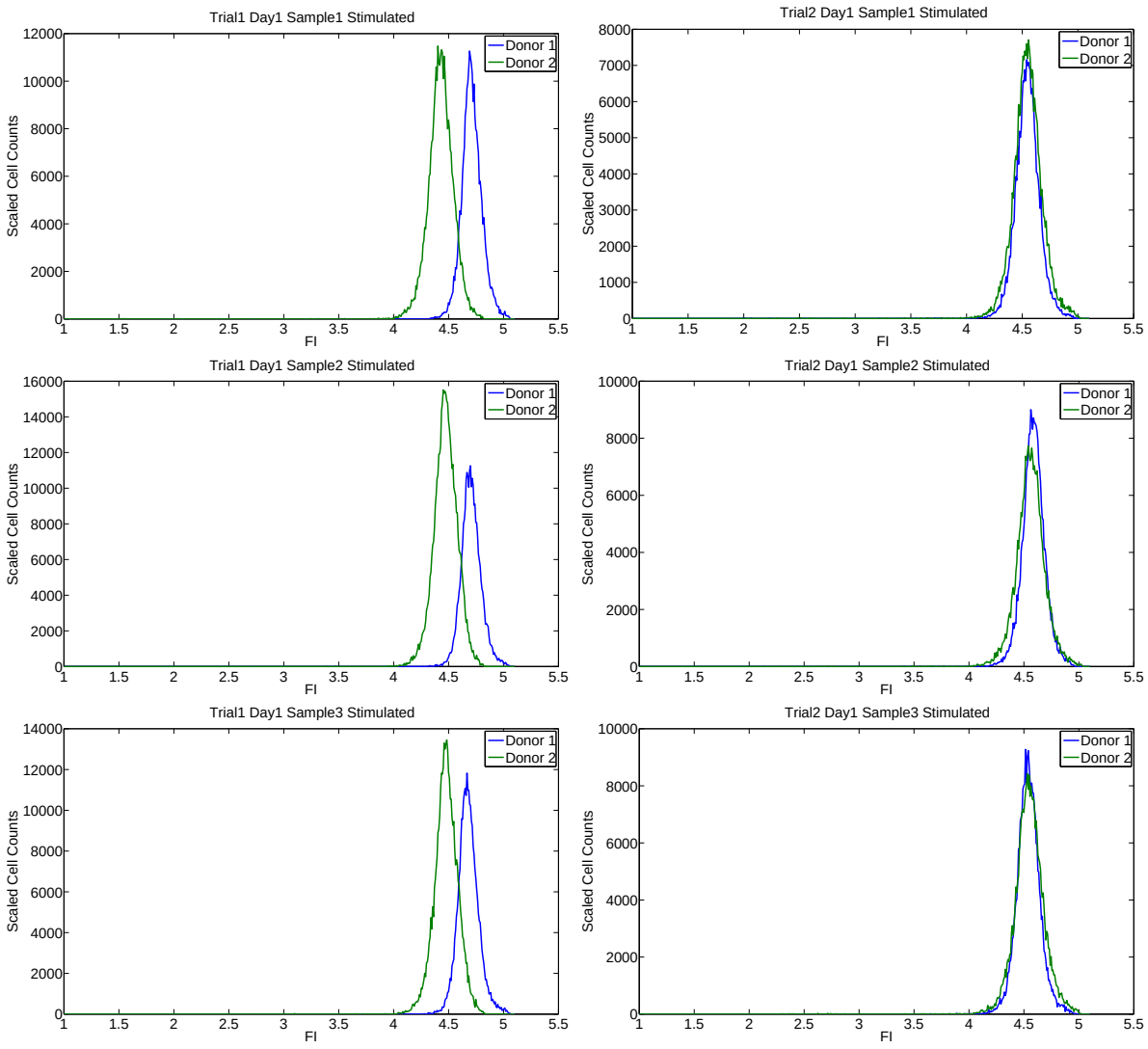

Figure 4.10: Inter-individual variability among cells stimulated with PHA at day 1. Left: Trial 1. Right: Trial 2. Again we find the cells from Donor 1 received a brighter initial staining than the cells from Donor 1 during Trial 1. The overall shape of the histograms (up to scaling and initial CFSE FI) is consistent for each trial and sample, which is an strong indication of consistent cell behavior.

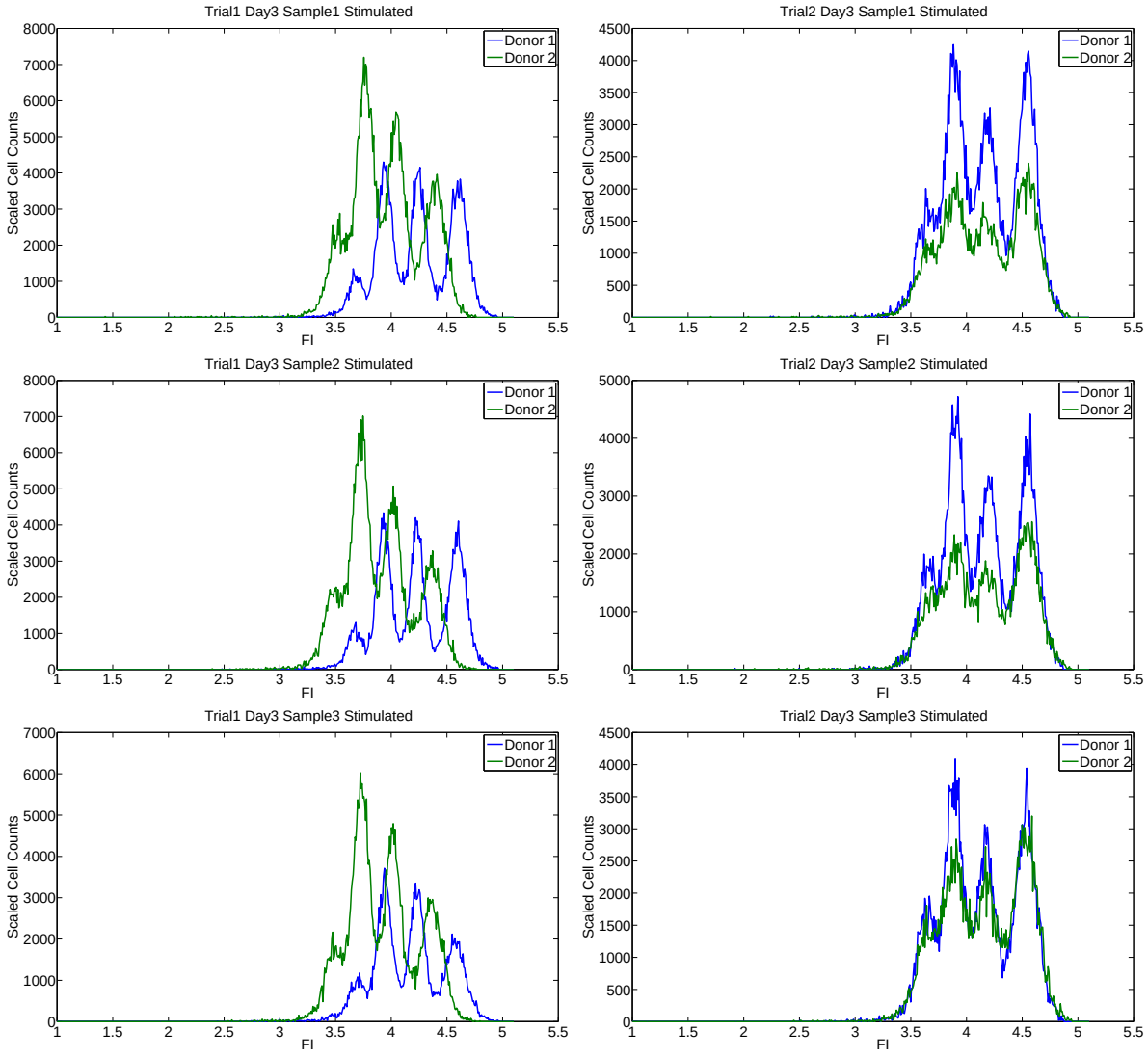

Figure 4.11: Inter-individual variability among cells stimulated with PHA at day 3. Left: Trial 1. Right: Trial 2. Again we find the cells from Donor 1 received a brighter initial staining than the cells from Donor 1 during Trial 1. The overall shape of the histograms (up to scaling and initial CFSE FI) is consistent for each trial and sample, which is an strong indication of consistent cell behavior.

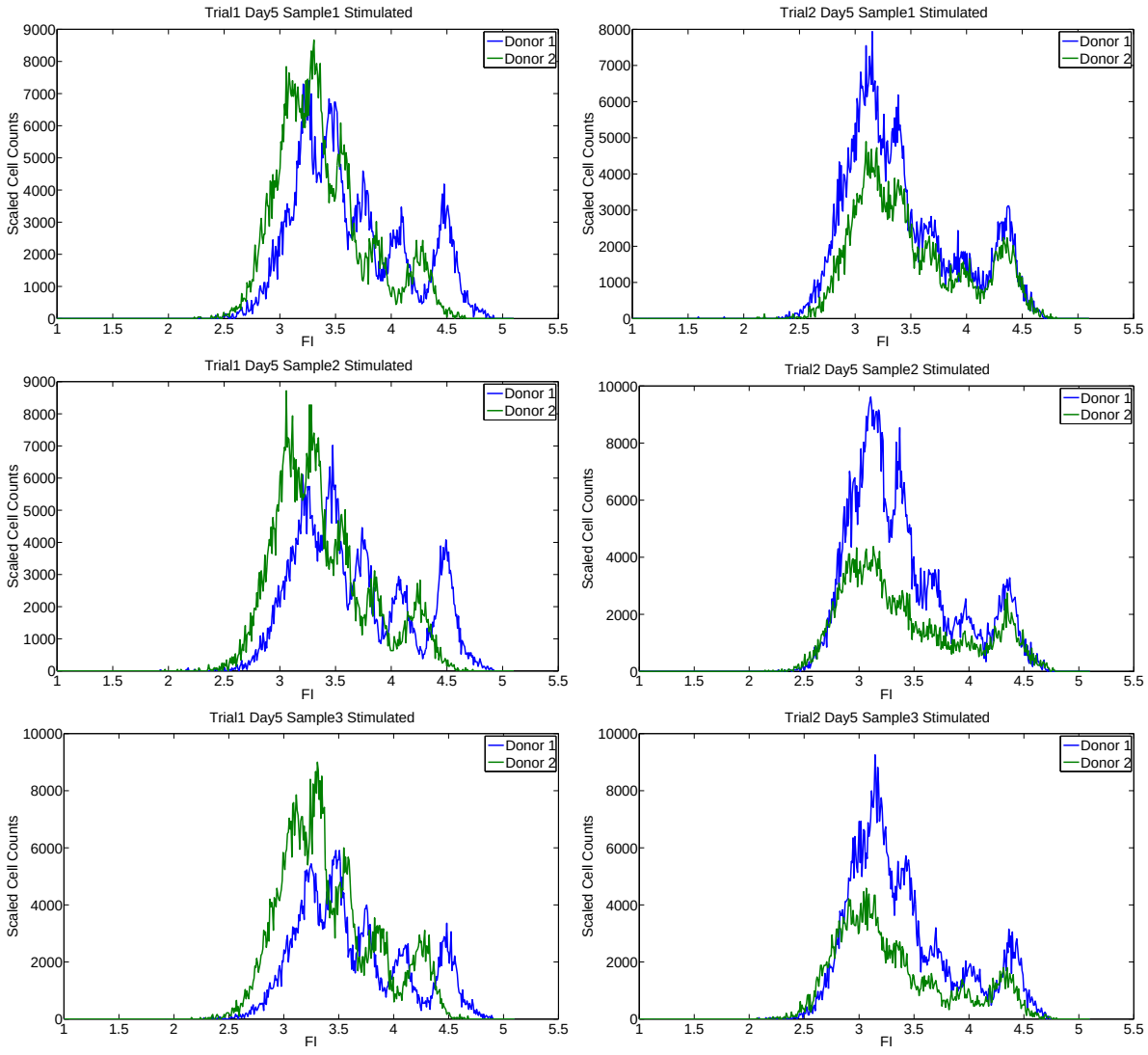

Figure 4.12: Inter-individual variability among cells stimulated with PHA at day 5. Left: Trial 1. Right: Trial 2.

given model to a data set, this may also provide a useful distinguishing feature to separate data from healthy and diseased donors. At the moment, such a claim is purely speculative–it will need to be verified by actually fitting a model (ideally, the compartmental model) to each data set and examining the degree to which the parameters of the model can be consistently and uniquely estimated. This will, in turn, almost certainly depend upon the times at which the data is taken. This work is a significant future goal.

4.4 Histogram Bins and Measurement Noise

So far, we have focused exclusively on experimental variability and intra- and inter-individual variability in CFSE histogram data. It has been deduced that the largest contributor to variability in CFSE histogram data is variability in the numbers of cells estimated to be in the population. Once unimportant differences (e.g., the exact time at which a measurement was taken, the total number of cells originally in placed in a well, etc.) in experimental setup are accounted for, the CFSE histogram data appears to be consistent for each donor from Trial 1 to Trial 2. Meanwhile, little attention has been paid to the actual statistical error model which accounts for the 'noise' in the histograms. From Figure 4.10 to Figure 4.12, notice that the level of noise in the data increases on each day. This would seem at first to indicate a time-dependence in the level of noise in the experimental data. However, for unstimulated cells (see Figures 4.5 and 4.6) the noise observed in the histogram data does not increase with time. We observe in Tables 4.4 and 4.5 that the number of PHA-stimulated cells actually measured by the flow cytometer (the number of cell events) decreases each day the cells are measured. Meanwhile, the number of beads counted decreases as well. In other words, a smaller sample (of the total population) is measured and then scaled up by a larger factor to form the population data. It is no surprise then, that the level of noise in the histogram data increases.

It is of interest to consider the effects of the histogram bins on the noise in the data. Thus far in this chapter, every histogram has been generated with 512 evenly spaced bins. (For the $CFSE^+$ cells, the bins are evenly spaced in the logarithmic scale. In Figure 4.13 three different numbers of bins (256, 512, and 1024) are used to form histograms for data collected from Donor 1 Trial 1 Day 5 (the data collected from Donor 2 is sufficiently similar; the effects of the histogram bins on noise in the data are the same for data from Trial 2 as well as days 1 and 3). For both donors and for all three measurements in the triplicate data sets, the level of noise (relative to the size of the histogram data) increases as the number of bins increases. This is as expected–when fewer bins are used, the bins themselves are larger and effectively smooth out the data. Given this observation, some work might be directed toward determining an optimal number of bins in which to place the flow cytometry data. If too few bins are used, significant features of the measured population may be overlooked. But if too many bins are used, the data will become too noisy to be accurately fit to the data.

The level of noise in the data can be reduced by averaging over the triplicate samples (weighted by the total number of cells estimated to be in the population for each sample). Averaged data for day 5 (both donors, both trials) can be found in Figure 4.14, where 512 bins have been used to construct the histograms. While there is less error in the averaged data than in any single one of the triplicate data sets (compare Figure 4.14 with the middle row of Figure 4.13), it is not clear that the decrease

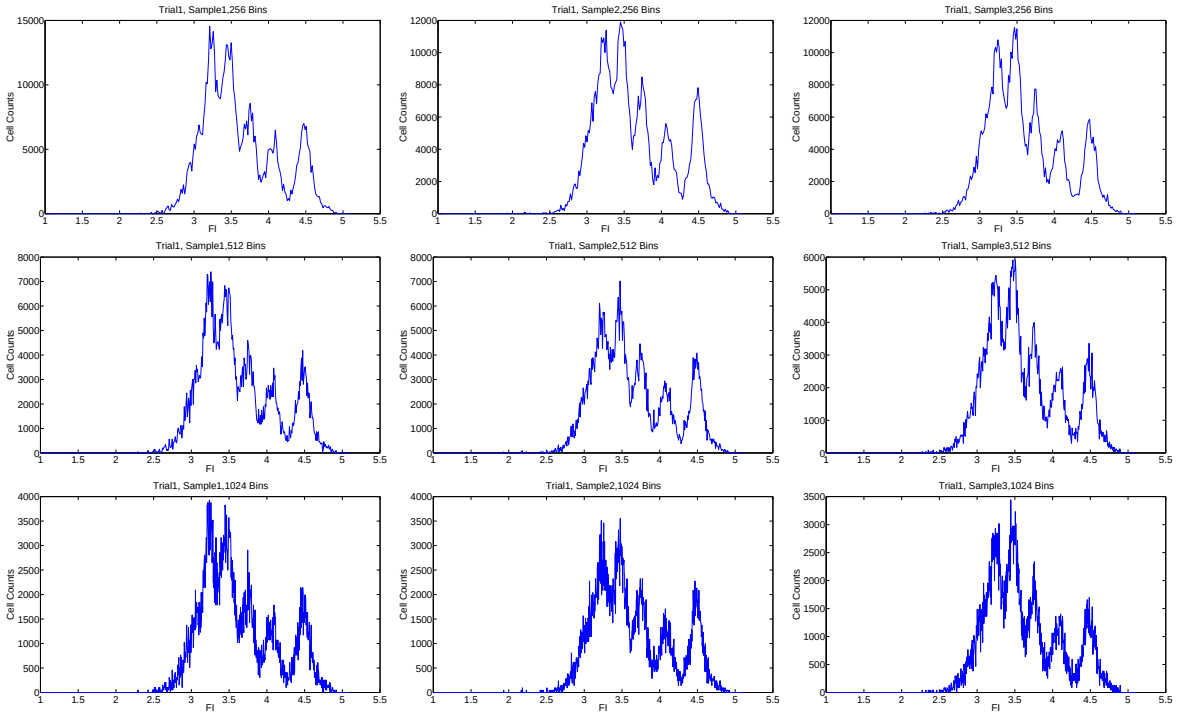

Figure 4.13: Effects of different numbers of bins on noise in the histogram data for Donor 1, Trial 1, day 5. Bins are evenly spaced in the logarithmic coordinate. Top: 256 bins. Middle: 512 bins. Bottom: 1024 bins. The effects are consistent for the triplicate data sets.

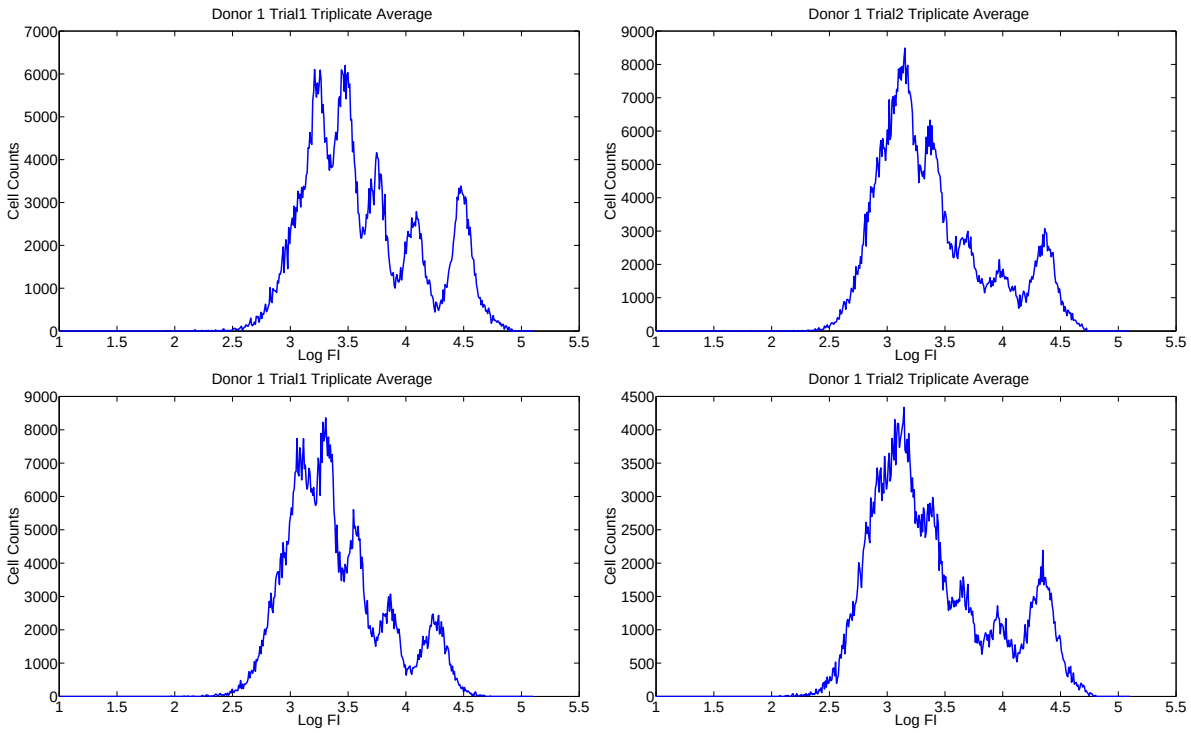

Figure 4.14: Averages of triplicate data for Donor 1 (top) and Donor 2 (bottom) on day 5. Left: Trial 1. Right: Trial 2.

in noise is sufficient to justify the technique. Rather than average across data sets, a similar effect could be obtained by taking only one measurement but measuring three times more cells. This seems advantageous, as it would have the same noise-reducing effect as averaging, but presumably with less risk of altering or biasing the results by combining multiple data sets.

We conclude this analysis by noting that the actual biophysical behavior of the populations of cells (as understood in terms of rates of division and death) seems very consistent from sample to sample within the same donor, and even from donor to donor among the two donors measured here. The primary source of variability between multiple data sets appears to be experimental (e.g., initial staining intensity, total number of cells initially placed in culture, etc.). For the actual dynamic parameters, the intra- and inter-individual variability seem quite small for healthy donors. This is encouraging, as it should allow for the set of admissible parameters (in the mathematical model) to be narrowed significantly, leading to faster optimization times. Moreover, if unhealthy donors (i.e., diseased donors) are described by parameters outside of this admissible range, this could provide a tool for distinguishing diseased from healthy cells. For the moment, this is purely speculation, and definitive proof relies upon the model being validated against additional data sets, as well as some additional mathematical work to demonstrate the uniqueness of the estimated parameters, etc. Still, these observations provide hope that the compartmental model can be combined with an accurate statistical model to establish biologically meaningful parameters as well as confidence bounds on those parameters.

4.5 Derivation of a Possible Statistical Error Model

Given the consistency of the flow cytometry data from the two donors, the most significant source of variability between different data sets appears to be experimental, involving the number of beads measured, number of cells counted, etc. This variability seems to be an inherent part of the data collection procedure and thus must be incorporated into the statistical model that links the mathematical model to the data. Recall the PDE model from Chapter 3:

$$\frac{\partial n_0}{\partial t} - ce^{-kt}(x - x_a)\frac{\partial n_0}{\partial x} = -(\alpha_0(t) + \beta_0(t) - ce^{-kt})n_0(t,x) \tag{4.3}$$

$$\frac{\partial n_1}{\partial t} - ce^{-kt}(x - x_a)\frac{\partial n_1}{\partial x} = -(\alpha_1(t) + \beta_1(t) - ce^{-kt})n_1(t,x) + R_1(t,x) \tag{4.4}$$

$$\vdots \tag{4.5}$$

$$\frac{\partial n_{i_{\max}}}{\partial t} - ce^{-kt}(x - x_a)\frac{\partial n_{i_{\max}}}{\partial x} = -(\beta_{i_{\max}}(t) - ce^{-kt})n_{i_{\max}}(t,x) + R_{i_{\max}}(t,x). \tag{4.6}$$

Here, $n_i(t,x)$ is the label-structured population density for cells of interest (e.g., CD4 cells) at time t with measured fluorescence intensity (FI) x having undergone i divisions. The parameters c and k describe the rate at which CFSE FI naturally decreases within a cell, and the parameter x_a is the natural AutoFI of cells. It follows that the structured population density for the total population is

$$n(t,x) = \sum_i n_i(t,x). \tag{4.7}$$

Finally, because CFSE histogram data is typically presented on a logarithmic scale, we can define the log FI variable $z = \log_{10}(x)$ and the new structured population density

$$\tilde{n}(t,z) = 10^z \ln(10) n(t, 10^z). \tag{4.8}$$

In order to derive an error model for the histogram data, we first make the common assumption that the structured population density $\tilde{n}(t,z)$ perfectly describes the population of cells. (Technically, this population density depends upon a set of parameters which must be estimated from the data; for simplicity we have suppressed this dependence. We are assuming that the model, once calibrated to data, represents 'truth'.) Then the total number of cells of interest in the population is

$$N(t) = \int_{x_a}^{\infty} \tilde{n}(t,z)dz. \tag{4.9}$$

Let t_j be the time at which a measurement is made. Previously, it has been assumed that such a measurement (after scaling by the bead ratio) represented a complete census of all relevant cells in the population. However, in the actual experimental procedure, we actually measure some number S_j of the cells. We can define the function

$$p_j(z) = \frac{\tilde{n}(t_j,z)}{N(t_j)}. \tag{4.10}$$

It follows that $p(z)$ is a probability density function. Thus, the sample of S_j cells (of interest) is taken

(without replacement) from the total population of $N(t_j)$ cells; the FI of the sampled cells is subject to the sampling density $p(z)$. (As addressed previously, and as should be carefully noted again here, there are numerous steps required to separate the cells of interest, e.g., CD4+ cells, from the actual culture of cells passing through the cytometer. References to the total number of cells $N(t)$, and the number of sampled cells S_j are understood to refer only to the specific cells of interest in the experiment. For the moment, we make the additional assumption that these two numbers are exact and are not subject to any errors–systematic, experimental, or otherwise–caused by gating, etc. These assumptions may need to be generalized, but we focus on the more simple case for the moment.) Let B be the total number of beads (51175 for the current data sets) which are used to quantify the fraction of the population of cells which is measured. Let b_j be the 'true' number of beads passing through the cytometer. (By this, we mean the exact number of beads which *would* pass through the cytometer if the measured culture were perfectly homogeneous, etc.). It follows that

$$S_j = \frac{b_j}{B} N(t_j). \tag{4.11}$$

Now consider the k^{th} histogram bin $[z_k, z_{k+1})$. The number of cells in the whole population which are contained in this bin is given by

$$I[\tilde{n}](t_j, z_k) = \int_{z_k}^{z_{k+1}} \tilde{n}(t_j, z) dz. \tag{4.12}$$

Let M_k^j be a random variable representing the number of cells counted into the k^{th} bin. (Note that M_k^j is a different random variable than the N_k^j used to describe the data in Chapter 3. M_k^j is the number of cells counted into a given bin *out of the cells actually passing through the flow cytometer*.) Because the measurement process represents a sampling without replacement, it follows that M_k^j is described by a hypergeometric distribution,

$$M_k^j \sim HypG(N(t_j), I[\tilde{n}](t_j, z_k), S_j). \tag{4.13}$$

Assume

- $N(t_j) >> S_j$
- $I[\tilde{n}(t_j, z_k)] >> \frac{S_j I[\tilde{n}](t_j, z_k)}{N(t_j)}$
- $0 < \epsilon \leq \frac{I[\tilde{n}(t_j, z_k)]}{N(t_j)} \leq 1 - \epsilon < 1.$

Then it can be shown [45] that $M_k^j \xrightarrow{distbn} \tilde{M}_k^j$ where

$$\begin{aligned}
\tilde{M}_k^j &\sim \mathcal{N}\left(\frac{S_j I[\tilde{n}](t_j, z_k)}{N(t_j)}, \frac{S_j I[\tilde{n}](t_j, z_k)}{N(t_j)}\left(1 - \frac{I[\tilde{n}](t_j, z_k)}{N(t_j)}\right)\right) \\
&= \mathcal{N}\left(\frac{b_j}{B}\frac{N(t_j) I[\tilde{n}](t_j, z_k)}{N(t_j)}, \frac{b_j}{B}\frac{N(t_j) I[\tilde{n}](t_j, z_k)}{N(t_j)}\left(1 - \frac{I[\tilde{n}](t_j, z_k)}{N(t_j)}\right)\right) \\
&\approx \mathcal{N}\left(\frac{b_j}{B} I[\tilde{n}](t_j, z_k), \frac{b_j}{B} I[\tilde{n}](t_j, z_k)\right).
\end{aligned} \tag{4.14}$$

Table 4.6: Summary of notation for the statistical model.

Notation	Description
$n(t,x)$	Original label structured density
$\tilde{n}(t,z)$	Log-transformed label structured density
$N(t)$	Total number of cells in the population at time t
$p_j(z)$	Probability density function from which cells are sampled
S_j	Number of cells sampled at time t_j
B	Total number of beads originally placed into each well
b_j	'True' number of beads counted by the cytometer at time t_j
$I[\tilde{n}](t_j,z_k)$	'True' number of cells from the total population belonging in the k^{th} histogram bin at time t_j
M_k^j	Random variable representing the number of cells counted into the k^{th} histogram bin at time t_j
$\hat{b}_j$	Actual number of beads counted by the flow cytometer (a realization of b_j)
N_k^j	Random variable resulting when M_k^j is scaled by the ratio $B/\hat{b}_j$
n_k^j	The actual data, a realization of N_j^k
λ_j	Random variable representing the bead count error ratio $b_j/\hat{b}_j$

The final approximation is valid provided $I[\tilde{n}](t_j,z_k) << N(t_j)$, which is a perfectly reasonable assumption. It can easily be shown that the first two assumptions above are satisfied if $b_j/B << 1$. We see from Tables 4.1 - 4.5 that $b_j/B < 0.08$ generally speaking, so that this assumption seems reasonable. It is somewhat paradoxical that the condition $b_j/B << 1$ implies only a very small sample of the total population is measured. The final assumption above bounds the probability that a cell belongs to a particular bin away from zero and one (although this assumption is not strictly necessary in some cases [70]). In practice, this assumption is only violated when $I[\tilde{n}(t_j,z_k)] \approx 0$.

Finally, when the measurements are actually taken, a certain number of beads $\hat{b}_j$ are actually counted. We would certainly hope that $\hat{b}_j \approx b_j$; however, we can think of $\hat{b}_j$ as a realization of some random variable (which may or may not be an unbiased estimator of b_j, per the discussions elsewhere in this document). Following the notation of Chapter 3, the histogram data (to which the compartmental model is actually fit) is the random variable

$$N_k^j = \frac{B}{\hat{b}_j} M_k^j \tag{4.15}$$

Thus,

$$\begin{aligned} N_k^j &\sim \mathcal{N}\left(\frac{B}{\hat{b}_j}\frac{b_j}{B} I[\tilde{n}](t_j,z_k), \frac{B^2}{\hat{b}_j^2}\frac{b_j}{B} I[\tilde{n}](t_j,z_k)\right) \\ &= \mathcal{N}\left(\lambda_j I[\tilde{n}](t_j,z_k), \lambda_j \frac{B}{\hat{b}_j} I[\tilde{n}](t_j,z_k)\right), \end{aligned} \tag{4.16}$$

where we have defined $\lambda_j = b_j/\hat{b}_j$.

This form for the statistical model of the data, in addition to having the advantage of the mathematical derivation presented in this section, also seems to explain several observations regarding CFSE

data, both in this chapter and in Chapter 3. First, it was assumed in Chapter 3 (and the assumption is common to other mathematical treatments of CFSE data as well) that

$$E[N_k^j] = I[\tilde{n}](t_j, z_k). \tag{4.17}$$

However, from Equation (4.16), we have $E[N_k^j] = E[\lambda_j]I[\tilde{n}](t_j, z_k)$. Unless $\lambda_j = 1$ for all j, then these two formulations are not equivalent. For a fixed j, if $\lambda_j \neq 1$, it follows that the model will be systematically in error when compared to the data. This may explain the nonrandom structured observed in Figure 4.3, as well as the positive correlation in Figure 4.4. As is seen in Tables 4.1 - 4.5 of this document, the values of $\hat{b}_j$ (and hence λ_j) can vary significantly (by as much as 40%) over the triplicate measurements. This may partially explain the sudden 'appearance' of additional cells in the data set. In other words, this new statistical model likely is capable of explaining the precursor cohort problem.

At the beginning of this chapter, residual plots were used to demonstrate that neither an Ordinary Least Squares (OLS) nor Generalized Least Squares (GLS) formulation for the error variance provides the appropriate randomness of the residuals. As discussed there, the OLS framework amounts to a tacit assumption that $Var_{OLS}(N_k^j) = \sigma_0^2$ while the GLS case assumes $Var_{GLS}(N_k^j) = \sigma_0^2(I[\tilde{n}](t_j, z_k))^2$. However, we can now see from Equation (4.16) that neither of these formulations contains the appropriate power of I. Let us assume (for simplicity) that $\lambda_j = 1$ for all j in Equation (4.16). Then we may rewrite the new statistical model (compare (3.27)) as

$$N_k^j = I[\tilde{n}](t_j, z_k)\left(1 + \frac{\mathcal{E}_{jk}}{\sqrt{I[\tilde{n}](t_j, z_k)}}\right), \tag{4.18}$$

where $\mathcal{E}_{jk} \sim \mathcal{N}(0, \sigma_j^2)$. We use σ_j to represent the quantity $B/\hat{b}_j$ as well as any other (unmodeled) contributions to the variance which may depend on the data collection time t_j but nothing else. As before, let the data n_j^k be realizations of the random variables N_k^j. It follows that the modified residuals

$$\frac{r_{jk}}{\sqrt{I[\tilde{n}](t_j, z_k)}} = \frac{I[\tilde{n}](t_j, z_k) - n_k^j}{\sqrt{I[\tilde{n}](t_j, z_k)}} \tag{4.19}$$

should be randomly distributed with mean zero and constant (for fixed t_j) variance. Notably, Nordon et al. [86] find heuristically a similar variance model for cell numbers. In order to test this assertion, both the residuals and modified residuals are shown in terms of the value of the model in Figure 4.15. While there are still several issues with this statistical model (for instance, the nonrandom structures observable in the residual plots, and the negative bias of the residuals when the model is small), the modified residuals (with the new statistical model) have generally constant variance for each measurement time. The nonrandom structures seem likely to be explained by the failure of the current mathematical and statistical model to account for the parameters λ_j in (4.16). Some difficulty also arises at $t = 24$ and 48 hours because of the small cohort of cell duplets which was not gated out of the population. (As noted in Chapter 1, such a cohort is not expected to be a problem in future data sets.)

The negative skew of the residuals for small model values can also be explained. Consider, for instance, the graphic for $t = 96$ hours in Figure 3.9, in the region $z \geq 3$. The model predicts few (if

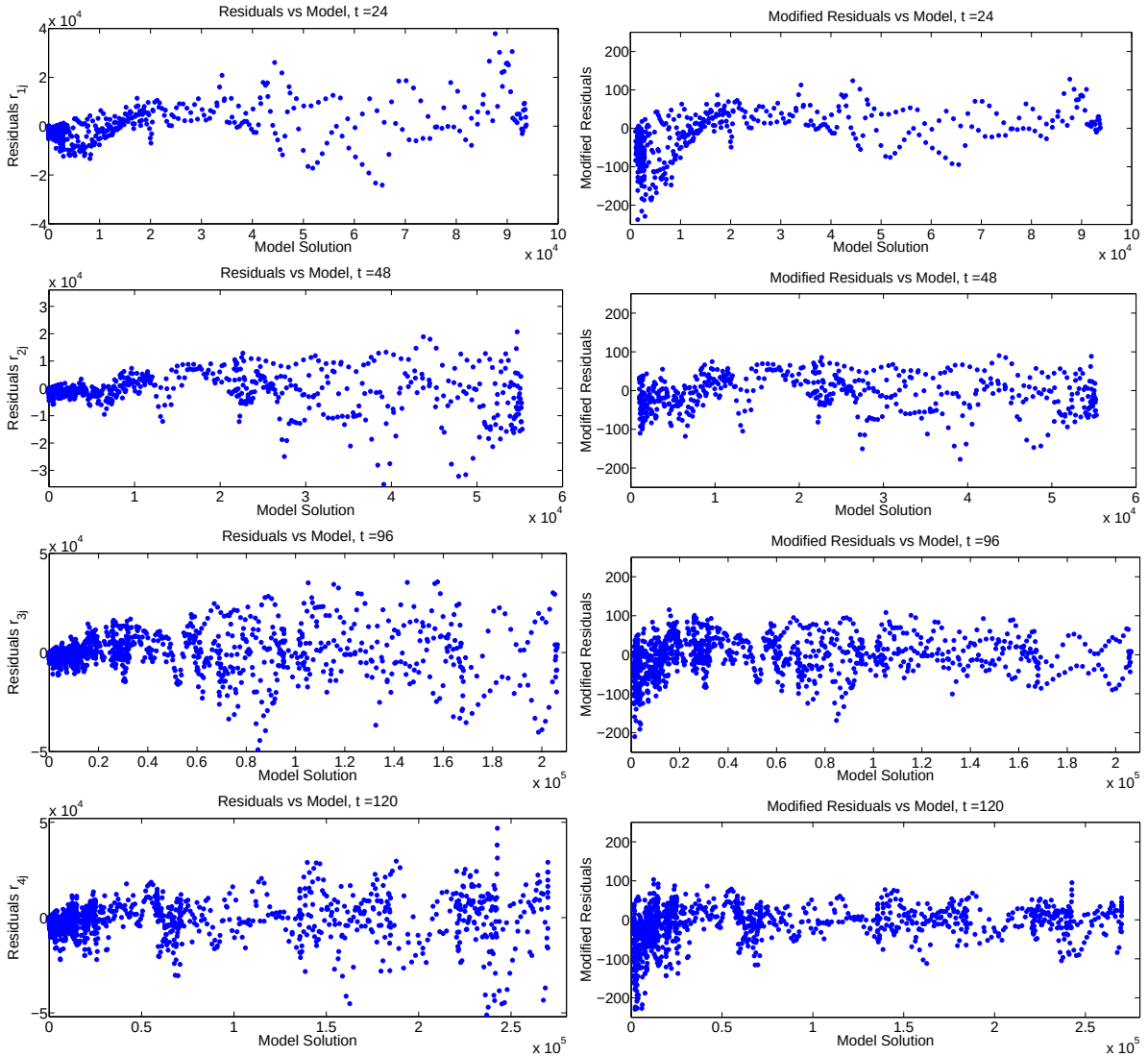

Figure 4.15: Left: Residuals r_{jk} plotted against the histogram model values $I[\tilde{n}](t_j, z_k)$. Right: Modified residuals $\frac{r_{jk}}{\sqrt{I[\tilde{n}](t_j,z_k)}}$ plotted against the same model values. In spite of the nonrandom patterns (most likely caused by the presence of cell duplets in the data) visible in the residual plots, the modified residuals appear to have a more constant variance.

any) cells in this region. While this prediction is accurate, we see that there is some noise in the data in this region. Because a histogram contains count data, these values must be positive, thus increasing the likelihood that the residual will be negative (because the value of the data exceeds that of the model). We can also see this from Equation (4.16). As $I[\tilde{n}](t_j, z_k) \to 0$, the assumption of normality must break down as the random variables N_k^j cannot attain negative values. This reflects a breakdown of one of the assumptions ($0 < \frac{I[\tilde{n}(t_j,z_k)]}{N(t_j)} < 1$) made in obtaining the new error model. This problem (and possible solutions) are considered briefly in Section 5.1.

The new statistical model seems to provide a significant improvement in relating the mathematical model from Chapter 3 to flow cytometry histogram data. When the sampling process used in collecting the data is explicitly accounted for, additional sets of previously unmodeled parameters $\{\lambda_j\}$ and $\{\hat{b}\}$ arise which can be used to explain both the apparent violation of a conservation law in some data sets as well as the observed variance in the error random variables. Some significant work is needed to examine how these additional parameters might be estimated (or, more specifically, to examine how the estimation of these parameters might affect the unique estimation of other parameters in the mathematical model). Once established, the new statistical model and the accompanying parameter estimation procedure will provide an important framework for the meaningful analysis of experimental data.

Chapter 5

Conclusions and Future Work

5.1 Implications of the New Statistical Model for Parameter Estimation

The error model derived at the end of the previous chapter seems promising in resolving several currently unexplained issues which have been noticed in flow cytometry data sets. First, by explicitly incorporating the effects of the scaling factor used in obtaining estimated population data from a sample of cells, the new error model (4.16) is capable of explaining the 'precursor cohort problem' and the apparent 'creation' of CFSE FI over the course of the experiment. Second, the new error model appears to accurately predict the variance of the 'noise' in the histogram cell counts. This analysis (Figure 4.15) must be taken with some skepticism, as the residuals used in the demonstration arise from a model which did not use any λ_j parameters to account for scaling of the data. The presence of cell duplets in the histogram data (particularly at $t = 24$ and 48 hours) also causes the accuracy of the compartmental model to break down slightly. These two features combine to result in residuals with nonrandom structure. Thankfully, in the newest CFSE data sets, it seems that the gating procedures used to collect the data have effectively eliminated any issues with cell duplets in the population; the new statistical model should resolve any additional discrepancies between the model and the data. Thus most (if not all) of the issues with nonrandom residuals highlighted in the previous chapter will not arise in future work.

It was observed in Figures 4.7 and 4.8 that the level of noise in the data seemed to increase as the ratio of beads counted to total beads decreased. This dependence appears directly in Equation (4.16). It is interesting to note that, while the level of noise expected in the data will decrease as $\hat{b}_j/B$ increases, the assumption of normality is premised upon the assumption $\hat{b}_j/B << 1$. Thus it may be possible to establish a desirable range of values for the number of counted beads in order to obtain optimal (in some sense) data from the experimental procedure.

At the moment, it is unclear how the new statistical model might be used most advantageously for the estimation of model parameters. Naively, one could use (4.16) to write

$$N_k^j = \lambda_j I[\tilde{n}](t_j, z_k) + \mathcal{E}_{kj}$$

where now the $\mathcal{E}_{kj}$ are random variables satisfying $E[\mathcal{E}_{kj}] = 0$ and $Var(\mathcal{E}_{kj}) = \lambda_j \frac{B}{\hat{b}_j} I[\tilde{n}](t_j, z_k)$. Thus, in addition to the model parameters θ (see Table 3.3), the parameters λ_j and $\hat{b}_j$ must be estimated as well. However, it seems almost certain that the addition of these new parameters will cause a troubling lack of identifiability when the model is fit to the data. For instance, in the absence of cell division, a decrease in total cell number from t_1 to t_2 is modeled identically by cell death (the parameters β_i) and by the scaling factors in the absence of any cell death (provided $\lambda_2 < \lambda_1$). As such, some additional work will be required to characterize the parameters λ_j and $\hat{b}_j$ and their dependence on the measurement time t_j. As more information is obtained regarding these parameters, it may be possible to establish a suitable framework in which they can be reliably and uniquely estimated.

An additional possibility for an inverse problem formulation with the new statistical model is likelihood estimation. Following the analysis of Section 4.5, Equation (4.16) provides an exact form (a normal distribution) for the likelihood of the cell count for a particular bin, rather than specifying only the mean and variance (as in a least squares framework). In the light of this additional information, it may possibly be advantageous to use the maximum likelihood principle for the estimation of a best-fit parameter vector. As in the least squares case, the estimation of the parameters λ_j and $\hat{b}_j$ may still be problematic. It may be possible (or even preferable) to use hierarchical modeling [37] to account for the hypothetical distributions from which the parameters λ_j and $\hat{b}_j$ are sampled. Again, more information regarding these parameters will need to be obtained from the experimental protocol before any such techniques can be reasonably used.

Regardless of the inverse problem framework chosen, some additional work is still needed to address the skew-negative residuals observed for small model values in Figure 4.15. This feature was partly explained in Section 4.5 by the breakdown of the assumptions required to obtain a normal distribution for the new error model. This may possibly be fixed in a likelihood framework by the direct use of the hypergeometric distribution to model the likelihood of a given cell count. However, because the data is count data and cannot obtain negative values, the residuals (model - data) must necessarily be negative when the model solution is at or near zero. It may be possible to fix this skew-negative problem by deliberately ignoring certain data points [18]. That is, if $I[\tilde{n}](t_j, z_k)$ is below some threshold value, then the data point (z_k^j, n_k^j) is dismissed as meaningless noise and is not considered in the target (least squares cost or likelihood) function. Some additional work and analysis would be necessary to justify such an assertion and to test its effects on the parameter estimation procedure.

5.2 Generalizations of the Statistical Model

The statistical model (4.16) was derived by considering a single histogram bin at a single time. The resulting statistical model (4.18) arises from repeating this derivation for each histogram bin at each measurement time. However, in this derivation, the dependence of the cell counts (and thus the probabilities) on additional factors has been completely ignored. The model values $I[\tilde{n}](t_j, z_k)$ will depend strongly on the set of bins $[z_k, z_{k+1})$ used for the histogram data. It was shown in Figure 4.13 that the level of noise in the data (relative to the magnitude of the data) increases as the number of bins increases. Conversely, as fewer bins are used, the data is effectively 'smoothed out' or averaged and some smaller

features of the population data may be lost. Thus there seems to be some optimal number of bins to use to represent the histogram data. On one hand, this possibility could be assessed by trial and error on the number of histogram bins. Alternatively, the statistical model might be analyzed and/or generalized to explicitly incorporate the dependence of the statistical model on the number of histogram bins, and more importantly, the dependence of parameter confidence intervals on the number of histogram bins.

There are several additional generalizations to consider. In the derivation of the new statistical model, it was tacitly assumed that the S_j cells sampled from the population of $N(t_j)$ cells are taken completely at random. In reality, the gating procedures needed to identify cells of interest from other elements of the PBMC culture will result in sources of error which may be nonrandom in some way (e.g., by selecting cells with certain characteristics). The use of histograms to represent a measured population of CFSE-labeled cells is mathematically convenient, but overlooks a large number of complex steps which are required to identify the cells which are to be studied. Even simple flow cytometer setups measure multiple properties of cells (such as size, granularity, and surface marker expression) in addition to CFSE FI which are then used to distinguish among different cell types to be studied. When gates are set to include (or exclude) a certain group of cells, this amounts to censoring the data (in a higher-dimensional data space); this data is then projected down to a single dimension (CFSE FI) and binned into the histograms which are used for model calibration. Ideally, the cells of interest in the PBMC culture can be easily identified and separated by the gating process. It is possible however, that some regions (in the higher dimensional data space) are harder to separate than others. This differential censoring of the data would then have effects of the statistical model. Significant additional work (with additional data sets) will be needed to examine the possibility of such effects.

Finally, the new statistical model (4.18) makes the simplifying assumption that the numbers of cells counted into each distinct histogram bin represents an independent (from the other bins) process. This is not true; if S_j cells are measured at time t_j, then we must have the identity $\sum_k M_k^j = S_j$. Thus the random variables representing the numbers of cells $N_k^j = \frac{B}{\hat{b}_j} M_k^j$ in the total population counted into distinct bins are not independent. At each measurement time, it may be possible to ignore the mutual dependence of the random variables M_k^j provided there is a significantly large number of bins, but this is not certain. Assuming the data points are not independent, the inverse problem framework (almost certainly in a likelihood setting) must be modified to account for correlation between the data points, and the accurate estimation of the covariance matrix will be necessary in order to establish confidence bounds on parameters.

Finally, it is assumed that the CFSE FI of each cell is measured perfectly. Yet it seems reasonable to assume that the measurement process is subject to some error. For a given cell, if z is the 'true' log FI, we actually measure $Z = z + \epsilon$, where ϵ is a random variable representing measurement error. Assuming ϵ is described by some probability distribution P, we would then define the new integral operator

$$\tilde{I}[\tilde{n}](t_j, z_k) = \int_{-\infty}^{\infty} \left(\int_{z_k - \epsilon}^{z_{k+1} - \epsilon} \tilde{n}(t_j, \zeta) d\zeta \right) dP(\epsilon).$$

The task would then lie in identifying the probability distribution describing the machine measurement error, and this new definition of $I[\tilde{n}]$ would be incorporated into the derivation in Section 4.5. Alternatively, one might consider deriving an population model similar to the ones presented in Chapters 2

and 3, but derived explicitly in terms of the mass of CFSE within the cells. In order to relate such an equation to the CFSE FI data, a constitutive relationship (e.g., an affine relationship FI = K(mass) + AutoFI) could be imposed relating the mass of CFSE to the measured fluorescence intensity. Following the analysis of Section 3.4.3, one could then consider a probability distribution on the autofluorescence parameter. Measurement error could also be incorporated by placing a probability distribution on the parameter K. It is unclear if such hierarchical modeling would have any advantages over the current modeling efforts.

5.3 Implications of the Statistical Model for Model Comparison

It has already been acknowledged that the possibility that the error terms are not independent (along with the fact that they are not homoscedastic, and possibly not normally distributed) prevents the unbiased computation of standard errors or confidence bounds for the estimated parameters [8, 22]. Perhaps just as significantly, the use of Akaike's Information Criterion in Chapter 3 to compare various parameterizations of the compartmental model was premised upon the assumptions of independent, homoscedastic, normally distributed data. As such, the AIC values reported in Table 3.5 must be taken with extreme caution, as the failure of the data to meet the necessary assumptions raises the possibility that two different models with similar OLS costs (e.g., parameterizations A5B4dist and A5B5dist) may be closer together (in the sense of the information theoretic relative Kullback-Leibler distance [31]) than Table 3.5 indicates.

In actual fact, the AIC is derived in a likelihood framework. Let $l_{kj}(\xi|\theta)$ be the likelihood (typically defined as a probability density function) of observing a particular value ξ at time t_j and bin z_k^j. Then the total likelihood of a given parameter θ is

$$\mathcal{L}(\theta|\{n_k^j\}) = \prod_{j,k} l_{kj}(n_k^j|\theta)$$

Of course, this likelihood depends upon the choice of model parameterization. This dependency is understood but generally ignored for the present discussion. Let $\hat{\theta}_{MLE}$ be the maximum likelihood estimate of θ (that is, the value of θ that maximizes $\mathcal{L}$) and assume $\log \mathcal{L}(\theta)$ is twice continuously differentiable. Then the AIC is defined [31] as

$$AIC = -2 \log \mathcal{L}(\hat{\theta}_{MLE}|\{n_k^j\}) + 2p, \tag{5.1}$$

provided the mathematical model is an accurate description of the observed data and a sufficiently large data sample has been collected [31, Ch. 7]. Unlike the AIC given in 3.4.5 for the least squares framework, Equation (5.1) is an accurate asymptotic estimate of information loss (provided the stated assumptions hold) regardless of the statistical properties of the model errors. As was noted in Section 3.4.5, Equation (3.35) is premised upon normally distributed errors with constant variance σ_0^2. In that case,

$$l_{kj}(\{n_k^j\}|\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(I[\tilde{n}](t_j, z_k^j) - n_k^j)^2}{2\sigma_0^2}\right)$$

and

$$\begin{aligned}\log \mathcal{L}(\hat{\theta}_{MLE}|\{n_k^j\}) &= \log\left[\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{MLE}}\right)\exp\left(-\frac{1}{2\hat{\sigma}^2_{MLE}}\sum_{j,k} r_{kj}^2\right)\right] \\ &= -\frac{m}{2}\log(2\pi) - \frac{m}{2}\log\left(\frac{J(\hat{\theta}_{OLS})}{m}\right) - \frac{m}{2},\end{aligned}$$

where m is the total number of data points. The latter equality relies upon the facts that $\hat{\theta}_{OLS} = \hat{\theta}_{MLE}$ and $\hat{\sigma}^2_{MLE} = J(\hat{\theta}_{OLS})/m$ for normally distributed, homoscedastic errors. We observe that the first and last terms above do not depend in any way upon the model parameterization. Because the AIC is measured on an interval scale (that is, the absolute magnitude of the AIC does not matter, only differences between AIC values) and the first and last terms are the same for all possible parameterizations, these terms can be ignored. Plugging the middle term into (5.1), we obtain the AIC for ordinary least squares (3.35). Thus we see that (5.1) is a more general form of (3.35) and can be used regardless of the exact form of the statistical model (provided one is willing to work in a likelihood setting). This may be particularly advantageous as it would not be necessary to approximate the hypergeometric statistical model by a normal distribution (see Section 4.5), although the large numbers involved may cause combinatorial problems. Alternatively, approximation by a normal distribution (albeit one with nonconstant variance) leaves hope that a similar analysis to that presented above may relate the AIC to a weighted least squares cost.

As noted above, it is not immediately clear whether there it is advantageous to work in a likelihood or least squares framework. What is clear is that a statistical model which assumes normally distributed, constant variance errors does not accurately model the statistical properties of the observed data. As such, the use of the AIC (3.35) may provide misleading results. In some cases, this is not expected to change the results of Chapter 3. For instance, it seems perfectly reasonable to conclude the necessity of time and division dependent proliferation as well as variable AutoFI, given the differences in OLS cost (Table 3.5). Beyond these features, however, several models (e.g., parameterizations A5B3dist, A5B4dist, and A5B5dist) have quite similar costs. Thus it appears necessary to revisit these possible parameterizations of the compartmental model an improved statistical model.

5.4 Generalizations of the Mathematical Model

Apart from issues involving the statistical model relating the mathematical model to the data, it has been shown that the compartmental model of Chapter 3 accurately reproduces the behavior of a PHA-stimulated population of CD4+ cells as represented in histogram data from a flow cytometry assay. This model accounts for the natural rate of CFSE FI decay resulting from turnover of the intracellular label as well as the autofluorescence of cells in the absence of any fluorescent labeling. Simple linear models are used to describe the rates of cell division and death.

At the moment, only a single CFSE data set has been examined and used to estimate the parameters of the mathematical model(s). It is believed that the compartmental model is quite general and should apply to a wide range of data sets from various experimental setups. Work is ongoing to collect additional

data sets to demonstrate such a wide applicability of the model. As additional data sets become available, several additional features may need to be considered at greater length. In this section, we consider some of these features which appear most immediate.

5.4.1 Generalizations of the Autofluorescence Parameter

Ultimately, it is hoped that the compartmental model can be generalized to account for multiple cell types both in vivo and in vitro. While the cells studied in this report were cultured in a saturating quantity of the stimulating agent PHA, cells in vivo (or even cells in vitro in a different experimental setup) will not experience such a strong, constant stimulation. As such, the possibility exists that some cells may return to a quiescent state during the proliferation assay. As discussed briefly in Section 2.2.1, it is known that the autofluorescence of a cell changes depending upon its state of activation. Using the additional data sets surveyed in Chapter 4, we investigate the intra-individual and inter-individual variability of cellular AutoFI and the manner in which it changes during activation.

It should be noted that because we are only interested in the shape of the AutoFI distribution (and not the absolute number of measured cells) no beads are used in the AutoFI samples to determine the total number of cells. The histogram data is normalized by the total number of measured cells to provide an indication of the shape of the AutoFI probability distribution. Only one sample of cells is measured for each trial, time, and donor (as opposed to the triplicate samples discussed in Chapter 4). All AutoFI figures are plotted in a linear scale rather than the logarithmic scale.

In order to assess the intra-individual variability of AutoFI measurements, we can examine differences in the measured AutoFI distributions for each donor and for each day between Trial 1 and Trial 2. These distributions are shown in Figures 5.1 and 5.2 for Donors 1 and 2, respectively. While we do find some consistency, there are subtle differences in the AutoFI distributions between the two trials. It seems unlikely that AutoFI (which is caused by the fluorescence properties of naturally occurring intracellular molecules) would change between the two trials as a result of any biological changes to the cells. These differences could be explained, however, by slight differences in cultural conditions or machine calibration between the two trials. This explanation is strengthened by the graphics in Figures 5.3 and 5.4. In these figures, the AutoFI distributions for the two donors are compared directly (unstimulated cells in Figure 5.3 and PHA-stimulated cells in Figure 5.4). We find that the distributions for the two donors are nearly identical. Thus, we must rule out biological and/or cultural variability as the cause of the discrepancies in Figures 5.1 and 5.2. This is also mathematically reassuring–all cells of a given type (and for a given machine calibration) have similar AutoFI properties, and this holds true for activated cells as well as unactivated cells.

Following the analysis above, we have the unusual (and apparently paradoxical) consequence that inter-individual variability is essentially negligible while intra-individual is not. We emphasize that this seems to be explained by variations in machine calibration from one day to the next. This would explain why the measured AutoFI distributions are slightly different from Trial 1 to Trial 2 for a single donor (but measured 17 days apart), while the data collected from two different donors (but measured in quick succession) are quite similar. This argument is further supported by a careful look at the measured AutoFI distributions for unstimulated cells. Because these cells never become activated, there should

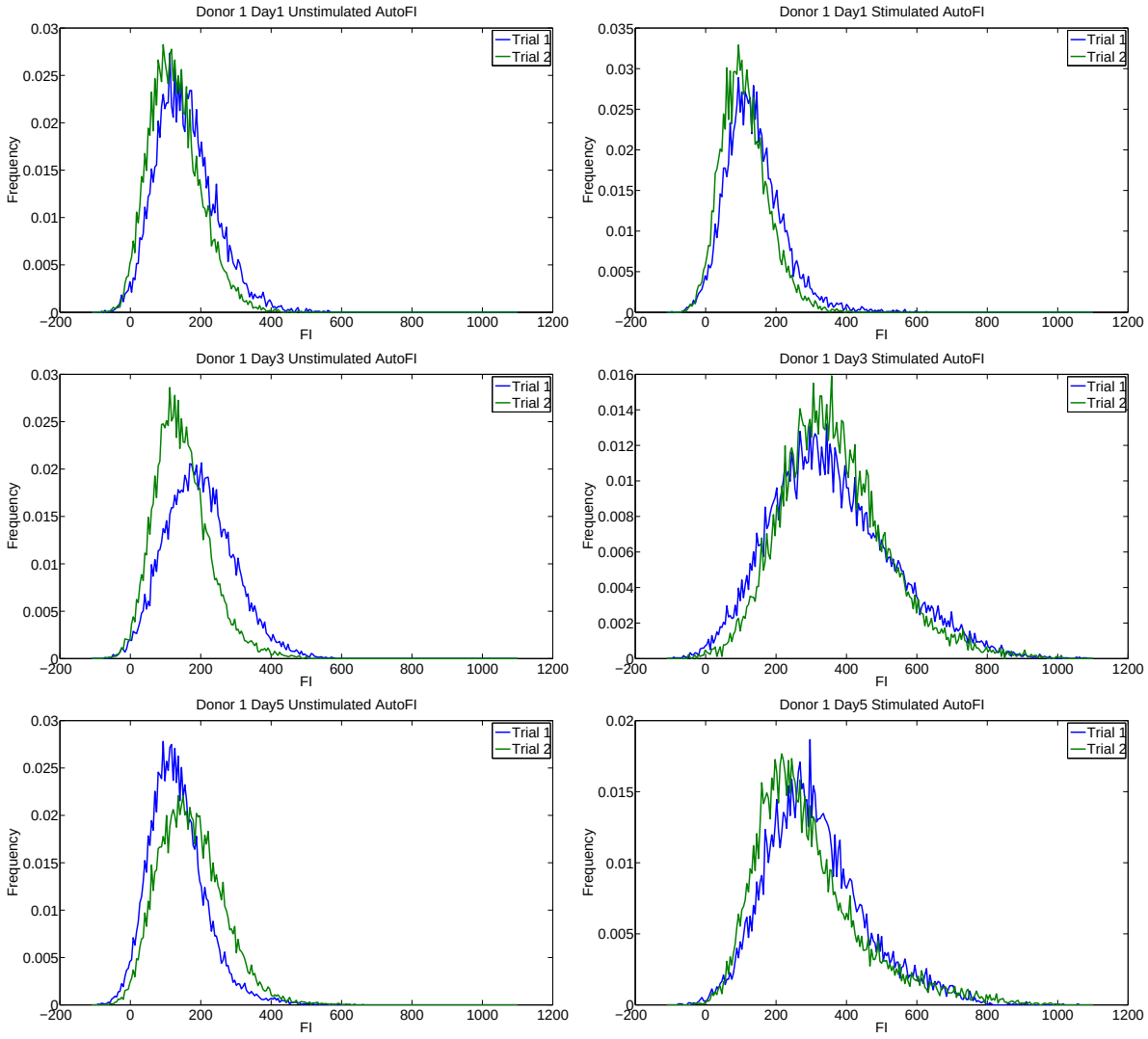

Figure 5.1: Data from Donor 1, showing intra-individual variability of AutoFI. Left: unstimulated cells. Right: cells stimulated with PHA. The changes between the two trials seem most likely to be the result of changes in machine calibration.

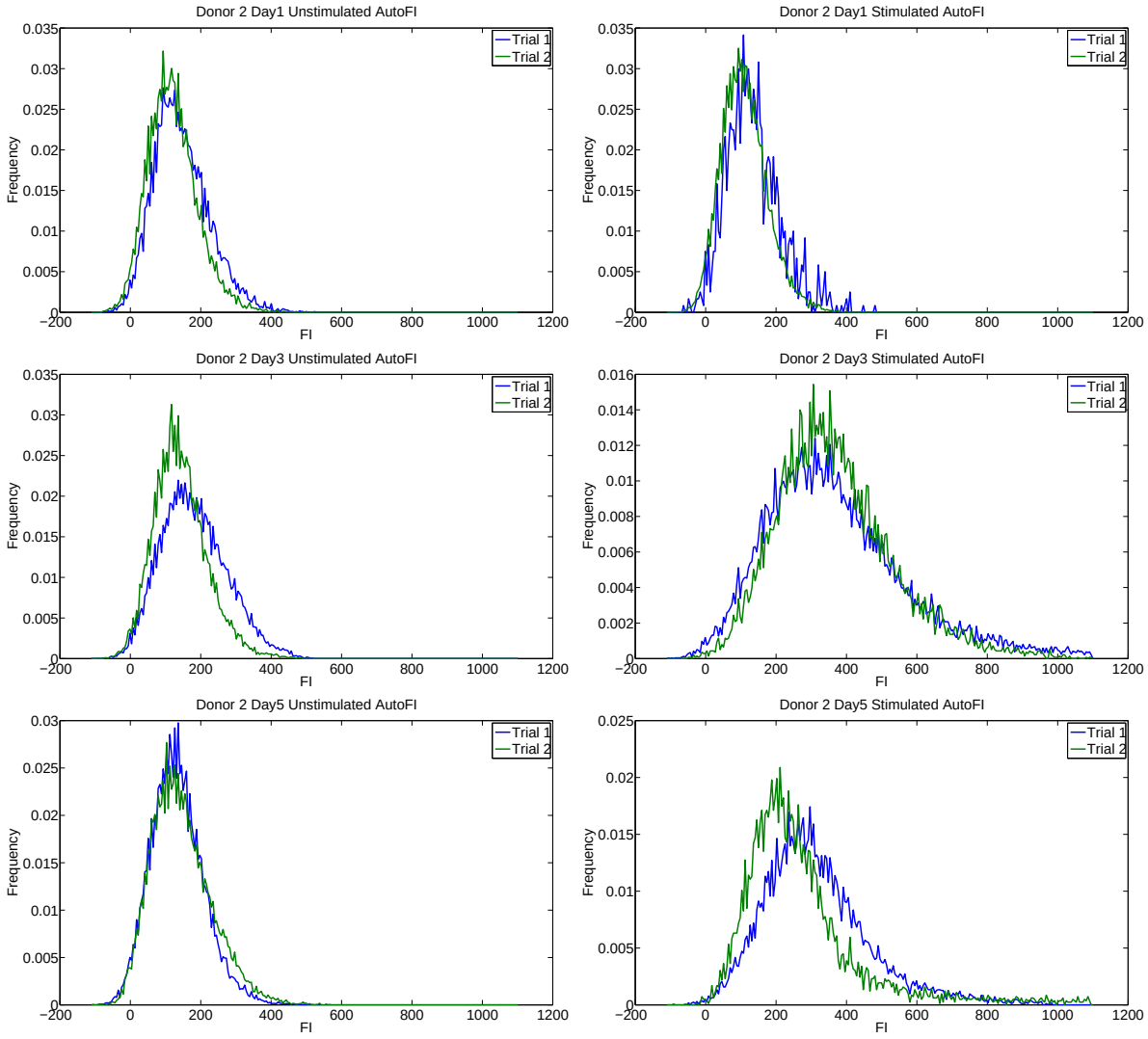

Figure 5.2: Data from Donor 2, showing intra-individual variability of AutoFI. Left: unstimulated cells. Right: cells stimulated with PHA. The changes between the two trials seem most likely to be the result of changes in machine calibration.

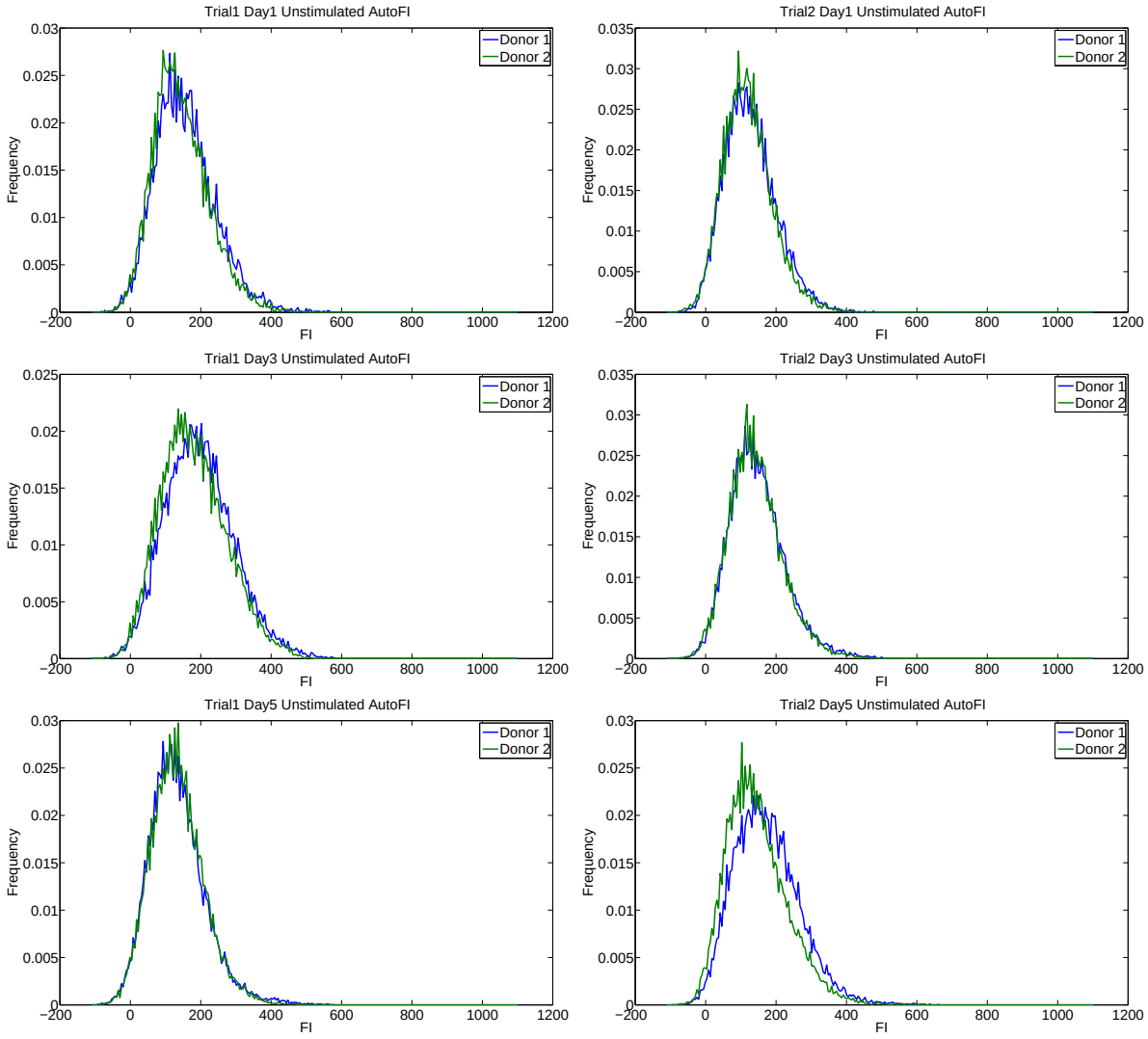

Figure 5.3: Inter-individual variability in AutoFI for unstimulated cells. Left: Trial 1. Right: Trial 2. Because the measurements were made in quick succession, there are no changes in machine calibration and inter-individual variability appears minimal.

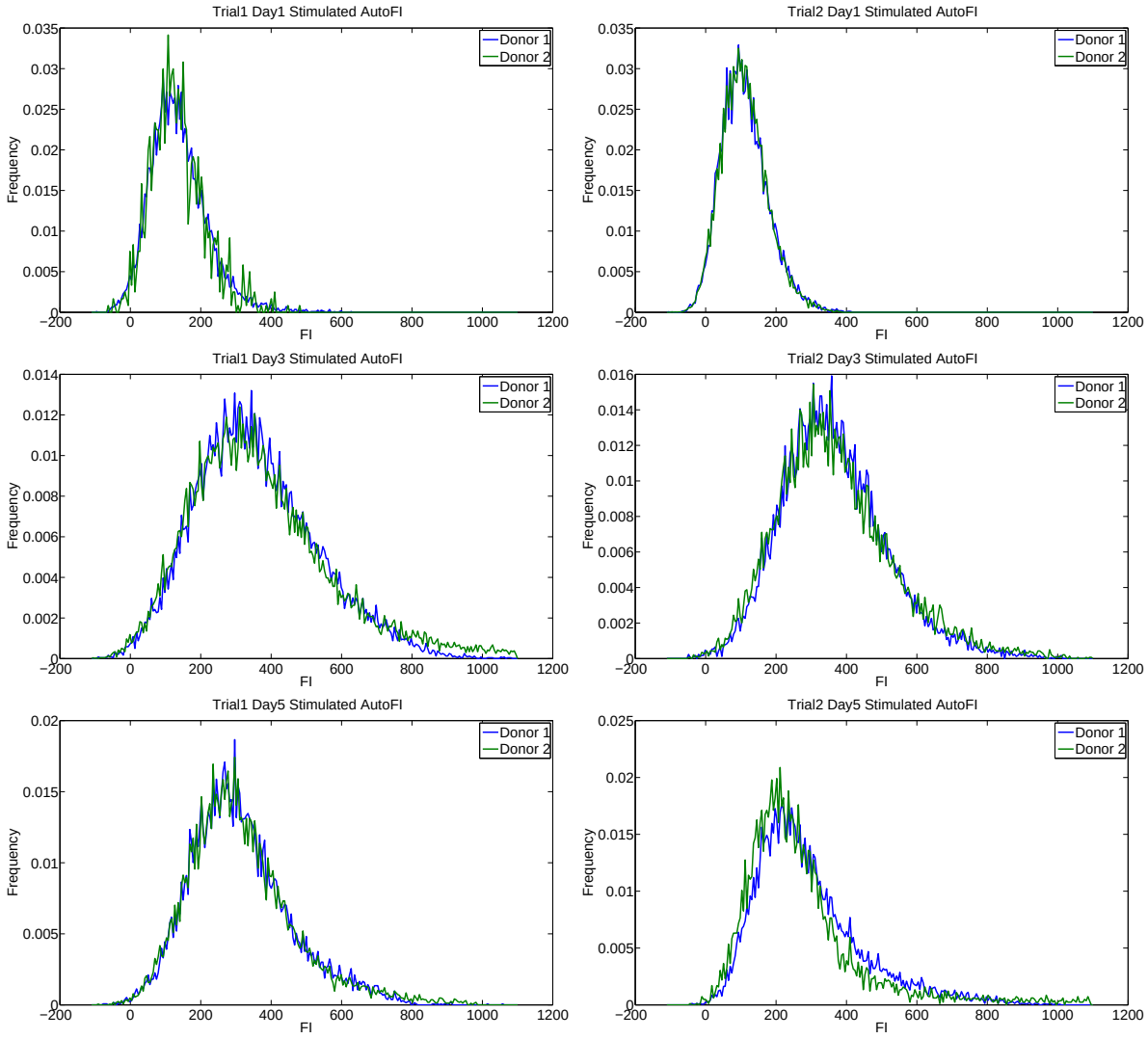

Figure 5.4: Inter-individual variability in AutoFI for cells stimulated with PHA. Left: Trial 1. Right: Trial 2. Because the measurements were made in quick succession, there are no changes in machine calibration and inter-individual variability appears minimal.

	Mean		Standard Deviation	
Data Set	Unstimulated	PHA-stimulated	Unstimulated	PHA-stimulated
Donor 1 Trial 1 Day 1	158.8729	139.5412	85.0167	82.3563
Donor 1 Trial 1 Day 3	199.4937	360.1652	96.8405	171.3622
Donor 1 Trial 1 Day 5	135.8948	314.5112	80.9387	141.9094
Donor 1 Trial 2 Day 1	129.6475	113.1297	73.7424	67.0125
Donor 1 Trial 2 Day 3	148.6522	366.5360	78.9271	148.9193
Donor 1 Trial 2 Day 5	175.0076	293.7099	92.0280	152.3741
Donor 2 Trial 1 Day 1	146.4155	137.4489	78.8734	77.1717
Donor 2 Trial 1 Day 3	182.9860	382.0809	93.1277	223.2650
Donor 2 Trial 1 Day 5	135.6401	317.2973	75.0909	152.5117
Donor 2 Trial 2 Day 1	122.0944	112.5239	69.9592	65.1638
Donor 2 Trial 2 Day 3	144.6255	369.9625	75.7685	164.6580
Donor 2 Trial 2 Day 5	147.5109	284.5370	85.5669	199.7716

Table 5.1: Summary of means and standard deviations for measured AutoFI distributions. Variability in the AutoFI distribution for unstimulated cells provides an estimate of the inherent variability of the experimental AutoFI measurements. While nonnegligble, this variability is small compared to the changes in AutoFI resulting from activation (compare unstimulated and PHA-stimulated AutoFI distributions).

be no major changes to their AutoFI properties. (This is in contrast to stimulated cells, as cells are known to increase in AutoFI when they become activated.) Yet, in Figure 5.3, the measured AutoFI distributions change slightly from day to day. For Trial 1, the distributions for days 1 and 5 are nearly identical, while the two distributions on day 3 have a larger mean and variance. For Trial 2, the mean and standard deviation increase each day. Because the day-to-day changes in AutoFI are different for the two different trials, these differences are more likely to be the result of machine calibration than actual biological changes.

The mean and standard deviation of each measured AutoFI distribution have been computed and these values are summarized in Table 5.1. Because the AutoFI distribution for unstimulated cells is not expected to change, the magnitude of variability associated with day-to-day changes in machine calibration (which are inherent in the experimental procedure) can be assessed by looking at the properties of the AutoFI distributions for unstimulated cells. We emphasize that such day-to-day variability is small in comparison to the changes in AutoFI associated with cell activation. In the current mathematical model, we have assumed that AutoFI can be accurately modeled either by its mean or possibly by a lognormal probability distribution (Chapters 3 and 4), but in both cases it is assumed that changes to AutoFI as a result of activation are negligible. Thus, if a more complex AutoFI mechanism is going to be considered for the mathematical model, it seems logical that activation-induced changes should be considered before any experimental variability. Future mathematical work (using either additional data sets or simulated data) will need to examine the effects of changing AutoFI on the model solution and determine whether or not such changes can be accurately identified in an inverse problem setting.

As noted in Chapter 2, activation-linked changes to AutoFI are believed to be negligible for the current experimental setup (because they only occur for undivided cells, and those cells have a fluorescence intensity due to CFSE which is significantly greater than their AutoFI). The reliability of this assump-

tion could be investigated mathematically if the compartmental for undivided cells were split into two compartments, based upon the state of activation of the cells. Each new compartmental would have its own AutoFI parameter (or distribution). Alternatively, it is possible to consider an AutoFI distribution which depends continuously on time (e.g., lognormal with mean $E[x_a](t)$ and standard deviation $STD(x_a)(t)$). Such time dependence is capable of incorporating the measurement-to-measurement variation in the AutoFI distributions, although this is likely to be highly data-set dependent and it is not clear if there is any need to consider such a feature in the model.

The best-fit parameterization of the compartmental model uses a lognormal density function to account for the variability of AutoFI in the population of cells. Alternatively, we could consider a nonparametric estimation of the probability distribution of AutoFI. Significant theory exists for such an estimation problem [4, 9, 12, 23], and it would be preferable to estimate the AutoFI distribution without imposing a particular functional form. Moreover, because the AutoFI distribution is directly observable, we have the possibility of comparing a nonparametric estimation of the distribution directly to the experimentally determined distribution. Such nonparametric estimation could be extended to encapsulate time-dependent behavior as well.

5.4.2 Generalizations of Proliferation and Death Rates

Following the results of [20, 21], we have used Malthusian rates for both proliferation (with time-dependent rates $\alpha_i(t)$) and death (with rates β_i). As discussed in Section 3.4, such an assumption is reasonable provided the turnover of cells (resulting either from division or death) occurs at a sufficiently rapid pace. Given the physiological constraints placed on rapidly dividing cells (e.g., rates of growth and DNA replication), one would expect some sort of minimum cell cycle time. It is unclear if the necessity of time dependence in the Malthusian rates α_i is an artifact of such a feature. To test this hypothesis, several generalizations of the proliferation and death rate terms are immediately available.

First, one might consider the addition of a second structure variable (say, volume) which could be used to enforce a minimum cell cycle time by requiring that cells progress from some size V to $2V$ before dividing, at which point two cells of size V are produced. However, in the absence of additional observations, it is unclear what parameters (e.g., average rate of growth, or the parameter V) might be estimable from CFSE histogram data. Video microscopy measurements by Hawkins et al. [58] indicate that average cell size may be division dependent, and this may add some additional complexity to the inclusion of volume structure. Biologically, it is expected that apoptosis occurs only at particular checkpoints in the cell cycle (particularly if external 'kill signals' are absent) so that a generalization to volume structure (or any other surrogate for cell cycle position or physiological age) may permit a more accurate description of cell death. Still, it is unclear what information might be estimated from only CFSE histogram data. It is possible that the forward scatter (FSC) of laser light may possibly be used as an observable surrogate for cell size, and some additional work will be necessary to investigate this possibility.

A second possibility to generalize the rates of proliferation and death would be to consider rate-limiting (e.g., logistic, Gompertz) models for proliferation and death. Some biological mechanisms have been proposed which may lead to density-dependent rates of cell death [33], and a Gompertz model for

cell growth has been used to account for quiescence in the context of a size-structured population model [54]. Of course, generalizations to nonlinear division and death must be considered in the context of the improvement they provide in fitting a given model to CFSE data sets. Given the accuracy of the simple linear models (albeit with time-dependent rates of proliferation), such generalizations seem unnecessary at the moment.

Given that the compartmental model (Chapter 3) can be used to compute numbers of cells per generation directly, some comparison has already been made between the results obtained with this model and the cell numbers computed from deconvolution techniques (Tables 3.7 and 3.8). It remains to compare the parameter estimates and model fits obtained with the compartmental model with those obtained from previous models (Smith-Martin, cyton, etc.). In fact, it maybe possible to incorporate into the compartmental model the mathematical forms used to describe proliferation and death in these models. Recall from Chapter 3 that the method of characteristics provides a solution (Equations 3.17 and 3.19) of the form

$$n_i(t, x(t;s)) = F(\text{Division,Death}) \tag{5.2}$$

where $x(t;s)$ is the characteristic line emanating from the point $(0, s)$ in the tx-plane. Clearly, the left side of Equation (5.2) is independent of any mathematization of cell proliferation and death. In Chapter 3, the form of the hypothetical function F is determined from the PDE formulation of the compartmental model (3.10) and the accompanying assumptions regarding the Malthusian rates of proliferation and death. Alternative, one could consider using (5.2) or its differential form (i.e., (3.18)) as a starting point, defining the right side of the equation in accordance with the assumptions of the Smith-Martin or cyton models, or their generalizations [42, 57, 71, 86, 106]. While previous authors have derived these models specifically in terms of total cell numbers, (5.2) could be related back to previous work by simple integration. The primary advantage in using (5.2) would be in the direct comparison of the model to histogram data, rather than from computed cell numbers. Further study could reveal the extent (if any) to which such a direct comparison improves the unique identification of parameters in previous models, although this will first rely on an accurate statistical model.

In this context, it is clear that several alternative possibilities exist for a mathematical description of proliferation and death rates. Thus it is clear that the interpretation of the proliferation and death parameters must be made with careful regard to the form of the model. Given the form of the model solution (Equations 3.17 and 3.19) for the compartmental model, it is plainly observed that linear changes in parameters for proliferation and death rates cause an exponential response in the computed solution [51, 40]. As such, the sensitivity of the model to these parameters, as well as the degree to which their estimation is unique, must be carefully considered when interpreting estimated parameters. The uniqueness of the estimated functions $\alpha_i(t)$ will depend on how the nodes for the linear splines are chosen in relation to the times at which data is taken. In some models, it has been shown that the effects of a linear increase of cell cycle time with division number cannot be distinguished from the effects of a linear increase in the death rate with division number [72]. If this is the case, then the biological interpretation of some parameters may be suspect.

Ideally, the values of $\alpha_i(t)$ and β_i can be related back to more physical/experimental parameters such as the type and strength of stimulation, which may in turn require the mathematization of certain

molecular pathways within individual cells. Recent work has indicated that the mechanisms responsible for cell proliferation and death may be mutually dependent upon a common molecular pathway [41, 100]. As more data becomes available, we hope to examine how the estimated parameters change under various experimental conditions, with an eye toward additional constitutive relationships linking molecular and/or subcellular functions to population dynamics [32]. In this context, it seems necessary to consider the extent to which these functions and/or pathways are inherited. Evidence suggests that closely related cells exhibit strong correlation in times to divide and some correlation in times to die, and that this correlation tends to decrease with the number of divisions undergone [58]. Cells with a common precursor may also share a common division destiny [58], which can be altered by stimulation conditions [102]. While computed cell numbers are relatively unaffected provided correlation is limited to cells having undergone the same number of divisions [42, 58, 62], correlation between subsequent division of cells can alter the dynamics predicted by a mathematical model [106]. For large populations, this effect seems negligible, but may play an important role in vivo where only a small number of responding cells can trigger an immune response [106]. As noted in the Chapter 1, cyton models and branching process models have been formulated to account for various levels of correlation, and these models may be incorporated into the compartmental model framework as described above. Alternatively, it may be possible (given any reasonable, identifiable parameterizations of cell division and death) to place probability distributions on these parameters (e.g., on the functions $\alpha_i(t)$ and $\beta_i(t)$) [6, 9, 12] in the manner described in Section 3.4.3.

5.4.3 Generalizations to Applications

In the research presented here, primary focus has been placed on the determination from data of biologically meaningful parameters which in some way describe the behavior of populations of cells, and by extension, can be used to describe the magnitude or efficacy of an immune response. Given additional data sets, the models presented to describe cell proliferation might be generalized to account for other division-dependent properties commonly observed in populations of cells. For instance, one might consider the emergence of differentiated subsets of cells as a population of cells responds to stimulation and divides [55, 86, 95]. Several authors have considered the role of quiescence in cellular populations [1, 53, 54], which may have implications for therapy/treatments [67]. The emergence of daughter cells might possibly be coupled with rates of mutation [65, 68, 108], which could have implications for drug resistance [66] or the emergence of tumors and their interaction with the environment [27]. Certainly, such work will depend upon additional experimental results and a mathematical and statistical model which can accurately and uniquely describe the dynamics of dividing cell populations, but such necessities do not appear unreasonable given the current modeling efforts. Given the general form and wide applicability of the model, it might be reasonably be used in a diagnostic setting [46] (e.g., to distinguish between healthy and diseased or abnormal cells based upon estimated proliferation rates).

5.5 Concluding Remarks

The compartmental model is the latest in a series of structured PDE models which can be fit directly to histogram representations of flow cytometry data. Once calibrated, the compartmental model can be used to quickly and accurately estimate the numbers of cells having undergone a certain number of divisions. This information can be used to determine biologically relevant parameters which will help to meaningfully compare cells from different donors and experiments. While the use of cell numbers per generation is not new, the direct modeling of histogram data reduces any need for deconvolution techniques which may introduce unnecessary bias into the computed cell numbers. Moreover, because the model is based upon conservation principles, it should be possible to fit histogram data even when the 'peaks' in the data (representing distinct generations of cells) are not well-resolved. This is a significant advantage over deconvolution techniques. The actual number of generations which can be accurately modeled (that is, the maximum value of $i_{\max}$) will depend upon the uniformity of the initial uptake of intracellular dye as well as the magnitude of the resulting CFSE FI relative to cellular AutoFI.

We are actively working to collect additional data sets with which to demonstrate the widespread applicability of this model, as well as to use this model in a systematic fashion to analyze how the estimated parameters vary under changing experimental and biological conditions. Most immediately, this will require the development of an accurate statistical model for the data. The generalization of the model to multiple cell types is immediate, although an accurate quantification of any interaction terms will require some careful thought and experimentation.

As more information becomes available regarding the complex processes involved in cell proliferation, we are confident that the model discussed here provides a firm physiological foundation upon which CFSE-based assay data can be understood. We strongly believe that the ideas and results presented here will form an important interpretive framework with a wide array of applications in experimental settings, diagnostic tests [46], and perhaps in a more integrated model of cell dynamics [63, 74].

REFERENCES

[1] O. Arino, E. Sanchez, and G.F. Webb, Necessary and sufficient conditions for asynchronous exponential growth in age structured cell populations with quiescence, *J. Mathematical Analysis and Applications*, **215** (1997), 499–513.

[2] B. Asquith, C. Debacq, A. Florins, N. Gillet, T. Sanchez-Alcaraz, A. Mosley, and L. Willems, Quantifying lymphocyte kinetics in vivo using carboxyfluorein diacetate succinimidyl ester, *Proc. R. Soc. B*, **273** (2006), 1165–1171.

[3] J.E. Aubin, Autoflouresecence of viable cultured mammalian cells, *J. Histochem. Cytochem.*, **27** (1979), 36–43.

[4] H.T. Banks and Kathleen Bihari, Modelling and estimating uncertainty in parameter estimation, *Inverse Problems*, **17** (2001), 95–111 .

[5] H. T. Banks, D. M. Bortz and S. E. Holte, Incorporation of variability into the modeling of viral delays in HIV infection dynamics, *Math Biosci.*, **183** (2003), 63–91.

[6] H.T. Banks, L.W. Botsford, F. Kappel, and C. Wang, Modeling and estimation in size structured population models, LCDS/CSS Report 87-13, Brown University, March 1987; *Proc. 2nd Course on Math. Ecology* (Trieste, December 8-12, 1986) World Scientific Press, Singapore, 1988, 521–541.

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

[8] H.T. Banks, M. Davidian, J. Samuels, and K.L. Sutton, An inverse problem statistical methodology summary, CRSC-TR08-01, North Carolina State University, January 2008; Chapter 11 in *Mathematical and Statistical Estimation Approaches in Epidemiology*, G. Chowell, et al., eds., Berlin Heidelberg New York, 2009, pp. 249–302.

[9] H.T. Banks and J.L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, *Appl. Num. Math.*, **57** (2007), 753–777.

[10] H.T. Banks and B.G. Fitzpatrick, Inverse problems for distributed systems: statistical tests and ANOVA, LCDS/CSS Report 88-16, Brown University, July 1988; *Proc. International Symposium on Math. Approaches to Envir. and Ecol. Problems*, Springer Lecture Notes in Biomath., **81** (1989), 262–273.

[11] H. T. Banks and B. G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, CAMS Tech. Rep. 89-4, Univ. of Southern California, September 1989; *J. Math. Biol.*, **28** (1990), 501–527.

[12] H.T. Banks and B.F. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, CAMS Tech. Rep. 90-2, Univ. of Southern California, January 1990; *Quart. Appl. Math.* **49** (1991), 215–235.

[13] H.T. Banks, K. Holm and D. Robbins, Standard error computations for uncertainty quantification in inverse problems: Asymptotic theory vs. Bootstrapping, CRSC-TR09-13, North Carolina State University, June 2009; Revised May 2010; *Mathematical and Computer Modelling*, **52** (2010), 1610–1625.

[14] H. T. Banks, Shuhua Hu and Zackary R. Kenz, A brief review of elasticity and viscoelasticity, CRSC-TR10-08, North Carolina State University, May 2010; *Advances in Applied Mathematics and Mechanics*, **3** (2011), 1–51.

[15] H.T. Banks and D.W. Iles, On compactness of admissible parameter sets: convergence and stability in inverse problems for distributed parameter systems, ICASE Report 86-38, NASA Langley Res. Ctr., Hampton, Virginia, 1986; *Proc. Conf. on Control Systems Governed by PDEs*, Gainesville, Florida. Springer Lecture Notes in Control and Inf. Sci., **97** (1987), 130–142.

[16] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

[17] H.T. Banks and M. Pedersen, Well-posedness of inverse problems for systems with time dependent parameters, CRSC-TR08-10, North Carolina State University, August 2008; *Arab. J. Sci. Eng. Math.* **1** (2009), 39–58.

[18] H. T. Banks and J. R. Samuels, Detection of cardiac occlusions using viscoelastic wave propagation, CRSC-TR08-23, North Carolina State University, December 2008; *Advances in Applied Mathematics and Mechanics,* **1** (2009), 1–28.

[19] H. T. Banks, R. C. Smith and Y. Wang, *Smart Material Structures: Modeling, Estimation and Control*, Masson Series on Research in Applied Math, Masson/J. Wiley, 1996.

[20] H.T. Banks, Karyn L. Sutton, W. Clayton Thompson, G. Bocharov, Marie Doumic, Tim Schenkel, Jordi Argilaguet, Sandra Giest, Cristina Peligero, and Andreas 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* (accepted).

[21] H.T. Banks, Karyn L. Sutton, W. Clayton Thompson, Gennady Bocharov, Dirk Roose, Tim Schenkel, and Andreas 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.

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

[23] H.T. Banks, B.G. Fitzpatrick, Laura K. Potter, and Yue Zhang, Estimation of probability distributions for individual parameters using aggregate population observations, CRSC-TR98-06, North Carolina State University, January 1998; *Stochastic Analysis, Control, Optimization, and Applications: a Volume in Honor of Wendell Fleming*, W. McEneaney, G. Yin, and Q. Zhang, eds, Birkhauser, Boston, 1999, 353–372.

[24] B. Basse, B. Baguley, E. Marshall, G. Wake, D. Wall, Modelling the flow cytometric data obtained from unperturbed human tumour cell lines: Parameter fitting and comparison, *Bull. Math. Biol.*, **67** (2005), 815–830.

[25] F. Bekkal Brikci, J. Clairambault, B. Ribba, and B. Perthame, An age-and-cyclin-structured cell population model for healthy and tumoral tissues, *J. Math. Biol.*, **57** (2008), 91–110.

[26] G. Bell and E. Anderson, Cell Growth and Division I. A Mathematical Model with Applications to Cell Volume Distributions in Mammalian Suspension Cultures, *Biophysical Journal*, **7** (1967), 329–351.

[27] N. Bellomo and L. Preziosi, Modelling and mathematical problems related to tumor evolution and its interation with the immune system, *Math. and Comp. Modeling*, **32** (2000), 413–452.

[28] S. Bernard, L. Pujo-Menjouet and M. C. Mackey, Analysis of cell kinetics using a cell division marker: Mathematical modeling of experimental data, *Biophysical Journal*, **84** (2003), 3414–3424.

[29] S. Bonhoeffer, H. Mohri, D. Ho, and A.S. Perelson, Quantification of cell turnover kinetics using 5-Bromo-2'-deoxyuridine, *J. Immunology*, **64** (2000), 5049–5054.

[30] Jose A. M. Borghans and R.J. de Boer, Quantification of T-cell dynamics: from telomeres to DNA labeling, *Immunological Reviews*, **216** (2007), 35–47.

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

[32] Nigel J. Burroughs and P. Anton van der Merwe, Stochasticity and spatial heterogeneity in T-cell activation, *Immunological Reviews*, **216** (2007), 69–80.

[33] Robin E. Callard, Jaroslav Stark, and Andrew J. Yates, Fratricide: a mechanism for T memory-cell homeostasis, *Trends in Immunology*, **24** (2003), 370–375.

[34] R. Callard and P.D. Hodgkin, Modeling T- and B-cell growth and differentiation, *Immunological Reviews*, **216** (2007), 119–129.

[35] R.J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, Chapman Hall, London, 2000.

[36] D. L. Chao, M. P. Davenport, S. Forrest and A. S. Perleson, Stochastic stage-structured modeling of the adaptive immune system, *Bioinformatics Conference CSB (2003): Proceedings 2003 IEEE*, Albuquerque, August 11-14, 124–131.

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

[38] R.J. DeBoer, V.V. Ganusov, D. Milutinovic, P.D. Hodgkin, and A.S. Perelson, Estimating lymphocyte division and death rates from CFSE data, *Bull. Math. Biol.*, **68** (2006), 1011–1031.

[39] R.J. DeBoer and Alan S. Perelson, Estimating division and death rates from CFSE data, *J. Comp. and Appl. Mathematics*, **184** (2005), 140–164.

[40] 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.

[41] Mark R Dowling, Dejan Milutinovic, and Philip D Hodgkin, Modelling cell lifespan and proliferation: is likelihood to die or to divide independent of age?, *J. R. Soc. Interface*, **2** (2005), 517–526.

[42] K. Duffy and V. Subramanian, On the impact of correlation between collaterally consanguineous cells on lymphocyte population dynamics, *J. Math. Biol.*, **59** (2009), 255–285.

[43] J.Z. Farkas, Stability conditions for the non-linear McKendrick equations, *Appl. Math. and Comp.*, **156** (2004), 771–777.

[44] J.Z. Farkas, Stability conditions for a non-linear size-structured model, *Nonlinear Analysis: Real World Applications*, **6** (2005), 962–969.

[45] W. Feller, *An introduction to probability theory and its applications, Volume I.* Wiley, New York, 1971.

[46] D.A. Fulcher and S.W.J. Wong, Carboxyfluorescein diacetate succinimidyl ester-based assays for assessment of T cell function in the diagnostic laboratory, *Immunology and Cell Biology*, **77** (1999), 559–564.

[47] Y. C. Fung, *A First Course in Continuum Mechanics*, Prentice Hall, Englewood Cliffs, NJ, 1994.

[48] Y. C. Fung, *Biomechanics: Mechanical Properties of Living Tissue*, Springer-Verlag, Berlin, 1993.

[49] Vitaly V. Ganusov, Dejan Milutinovi, and Rob J. De Boer, IL-2 regulates expansion of CD4+ T cell populations by affecting cell death: insights from modeling CFSE data, *J. Immunology*, **179** (2007), 950–957.

[50] V.V. Ganusov, S.S. Pilyugin, R.J. De Boer, K. Murali-Krishna, R. Ahmed, and R. Antia, Quantifying cell turnover using CFSE data, *J. Immunological Methods*, **298** (2005), 183–200.

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

[52] B.F. de St. Groth, A.L. Smith, W. Koh, L. Girgis, M. Cook, P. Bertolino, Carboxyfluorescein diacetate succinimidyl ester and the virgin lymphocyte: a marriage made in heaven, *Immunology and Cell Biology*, **77** (1999), 530–538.

[53] M. Gyllenberg and G. F. Webb, Age-size structure in populations with quiescence, *Mathematical Biosciences*, **86** (1987), 67–95.

[54] M. Gyllenberg and G. F. Webb, A nonlinear structured population model of tumor growth with quiescence, *J. Math. Biol.*, **28** (1990), 671–694.

[55] J. Hasbold, A.V. Gett, J.S. Rush, E. Deenick, D. Avery, J. Jun, and P.D. Hodgkin, Quantitative analysis of lymphocyte proliferation and differentiation in vitro using carboxyfluorescein diacetate succinimidyl ester, *Immunology and Cell Biology*, **77** (1999), 516–522.

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

[57] 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.

[58] E.D. Hawkins, J.F. Markham, L.P. McGuinness, and P.D. Hodgkin, A single-cell pedigree analysis of alternative stochastic lymphocyte fates, *Proc. Natl. Acad. Sci*, **106** (2009), 13457–13462.

[59] P.D. Hodgkin, J. Lee, A.B. Lyons, B cell differentiation and isotype switching is related to division cycle number, *J. Exp. Med.*, **184** (1996), 277–281.

[60] Mirja Hommel and Philip D. Hodgkin, TCR affinity promotes CD8+ T-cell expansion by regulating survival, *J. Immunology*, **179** (2007), 2250–2260.

[61] O. Hyrien and M.S. Zand, A mixture model with dependent observations for the analysis of CFSE-labeling experiments, *J. American Statistical Association*, **103** (2008), 222–239.

[62] O. Hyrien, R. Chen, and M.S. Zand, An age-dependent branching process model for the analysis of CFSE-labeling experiments, *Biology Direct*, **5** (2010), Published Online.

[63] D.E. Kirschner, S.T. Chang, T.W. Riggs, N. Perry, and J.J. Linderman, Toward a multiscale model of antigen presentation in immunity, *Immunological Reviews*, **216** (2007), 93–118.

[64] Kap-Hyoun Ko, Ross Odell, and Robert E. Nordon, Analysis of cell differentiation by division tracking cytometry, *Cytometry Part A*, **71** (2007), 773–782.

[65] Natalia L. Komarovaa, Anirvan Senguptac, and Martin A. Nowak, Mutationselection networks of cancer initiation: tumor suppressor genes and chromosomal instability, *J. Theoretical Biology*, **223** (2003), 433–450.

[66] N.L. Komarova, Stochastic modeling of drug resistance in cancer, *J. Theoretical Biol.*, **239** (2006), 351–366.

[67] N.L. Komarova and D. Wodarz, Effect of cellular quiescence on the success of targeted CML therapy, *PloS ONE*, **2** (2007), e990.

[68] Natalia L. Komarova, Lin Wu, and Pierre Baldi, The fixed-size LuriaDelbruck model with a nonzero death rate, *Mathematical Biosciences*, **210** (2007), 253–290.

[69] M. Kot, *Elements of Mathematical Ecology*, Cambridge UP: Cambridge, UK, 2001.

[70] S.N. Lahiri, A. Chatterjee, and T. Maiti, Normal approximation to the hypergeometric distribution in nonstandard cases and a sub-Gaussian Berry-Esseen theorem, *Journal of Statistical Planning and Inference*, **137** (2007), 3570-3590.

[71] H.Y. Lee, E.D. Hawkins, M.S. Zand, T. Mosmann, H. Wu, P.D. Hodgkin, and A.S. Perelson, Interpreting CFSE obtained division histories of B cells in vitro with Smith-Martin and Cyton type models, *Bull. Math. Biol.*, **71** (2009), 1649–1670.

[72] H.Y. Lee and A.S. Perelson, Modeling T cell proliferation and death in vitro based on labeling data: generalizations of the Smith-Martin cell cycle model, *Bull. Math. Biol.*, **70** (2008), 21–44.

[73] K. Leon, J. Faro, and J. Carneiro, A general mathematical framework to model generation structure in a population of asynchronously dividing cells, *J. Theoretical Biology*, **229** (2004), 455–476.

[74] Y. Louzoun, The evolution of mathematical immunology, *Immunological Reviews*, **216** (2007), 9–20.

[75] 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.

[76] T. Luzyanina, M. Mrusek, J.T. Edwards, D. Roose, S. Ehl, and G. Bocharov, Computational analysis of CFSE proliferation assay, *J. Math. Biol.*, **54** (2007), 57–89.

[77] 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.

[78] A. B. Lyons, Divided we stand: tracking cell proliferation with carboxyfluorescein diacetate succinimidyl ester, *Immunology and Cell Biology*, **77** (1999), 509–515.

[79] A. B. Lyons, Analysing cell division in vivo and in vitro using flow cytometric measurement of CFSE dye dilution, *J. Immunological Methods*, **243** (2000), 147–154.

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

[81] A. B. Lyons and K. V. Doherty, Flow cytometric analysis of cell division by dye dilution, *Current Protocols in Cytometry*, (2004), 9.11.1-9.11.10.

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

[83] J. E. Marsden and T. J. R. Hughes, *Mathematical Foundations of Elasticity*, Dover Publications, Inc., Mineola, NY, 1994.

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

[85] J.A. Metz and O. Diekmann, *The Dynamics of Physiologically Structured Populations*, Springer Lecture Notes in Biomathematics **68**, Heidelberg, 1986.

[86] Robert E. Nordon, Kap-Hyoun Ko, Ross Odell, and Timm Schroeder, Multi-type branching models to describe cell differentiation programs, *J. Theoretical Biology*, **277** (2011), 7–18.

[87] 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.

[88] R.W. Ogden, *Non-Linear Elastic Deformations*, Dover Publications, Inc., Mineola, NY, 1984.

[89] C. Parish, Fluorescent dyes for lymphocyte migration and proliferation studies, *Immunology and Cell Biol.*, **77** (1999), 499–508.

[90] B. Perthame, *Transport Equations in Biology,* Birkhauser Frontiers in Mathematics, Basel, 2007.

[91] Sergei S. Pilyugina, Vitaly V. Ganusov, Kaja Murali-Krishnac, Rafi Ahmed, and Rustom Antia, The rescaling method for quantifying the turnover of cell populations, *J. Theoretical Biology*, **225** (2003), 275–283.

[92] 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.

[93] P. Revy, M. Sospedra, B. Barbour, and A. Trautmann, Functional antigen-independent synapses formed between T cells and dendritic cells, *Nature Immunology*, **2** (2001), 925–931.

[94] I. W. Sandberg, Global implicit function theorems, *IEEE Trans Circuits and Systems*, **CAS-28** (1981), 145–149.

[95] T.E. Schlub, V. Venturi, K. Kedzierska, C. Wellard, P. Doherty, S.J. Turner, R.M. Ribeiro, P.D. Hodgkin, and M.P. Davenport, Division-linked differentiation can account for CD8+ T-cell phenotype in vivo, *Eur. J. Immunology*, **39** (2009), 67–77.

[96] G.A. Sever and C.J. Wild, *Nonlinear Regression*, Wiley, Hoboken, 2003.

[97] J. Sinko and W. Streifer, A New Model for Age-Size Structure of a Population, *Ecology*, **48** (1967), 910–918.

[98] J.A. Smith and L. Martin, Do Cells Cycle?, *Proc. Natl. Acad. Sci*, **70** (1973), 1263–1267.

[99] V.G. Subramanian, K.R. Duffy, M.L. Turner and P.D. Hodgkin, Determining the expected variability of immune responses using the cyton model, *J. Math. Biol.*, **56** (2008), 861–892.

[100] David T. Terrano, Meenakshi Upreti and Timothy C. Chambers, Cyclin-dependent kinase 1-mediated Bcl-x $_L$/Bcl-2 phosphorylation acts as a functional link coupling mitotic arrest and apoptosis, *Mol. Cell. Biol.*, **30** (2010), 640–656.

[101] B. Tummers, DataThief III. 2006 (http://www.datathief.org/)

[102] M.L. Turner, E.D. Hawkins, and P.D. Hodgkin, Quantitative regulation of B cell division destiny by signal strength, *J. Immunology*, **181** (2008), 374–382.

[103] H. Veiga-Fernandez, U. Walter, C. Bourgeois, A. McLean, and B. Rocha, Response of naive and memory CD8+ T cells to antigen stimulation in vivo, *Nature Immunology*, **1** (2000), 47–53.

[104] P.K. Wallace, J.D. Tario, Jr., J.L. Fisher, S.S. Wallace, M.S. Ernstoff, and K.A. Muirhead, Tracking antigen-driven responses by flow cytometry: monitoring proliferation by dye dilution, *Cytometry A*, **73** (2008), 1019–1034.

[105] Hilary S. Warren, Using carboxyfluorescein diacetate succinimidyl ester to monitor human NK cell division: Analysis of the effect of activating and inhibitory class I MHC receptors, *Immunology and Cell Biology*, **77** (1999), 544–551.

[106] C. Wellard, J. Markham, E.D. Hawkins, and P.D. Hodgkin, The effect of correlations on the population dynamics of lymphocytes, *J. Theoretical Biology*, **264** (2010), 443–449.

[107] J.M. Witkowski, Advanced application of CFSE for cellular tracking, *Current Protocols in Cytometry*, (2008), 9.25.1–9.25.8.

[108] Xiaoping Xiong, James M. Boyett, Robert G. Webster, and Juergen Stech, A stochastic model for estimation of mutation rates in multiple-replication proliferation processes, *J. Math. Biol.*, **59** (2009), 175–191.

[109] A. Yates, C. Chan, J. Strid, S. Moon, R. Callard, A.J.T. George, and J. Stark, Reconstruction of cell population dynamics using CFSE, *BMC Bioinformatics*, **8** (2007), Published Online.