Published in final edited form as:

J Immunol Methods. 2011 October 28; 373(1-2): 143–160. doi:10.1016/j.jim.2011.08.014.

A New Model for the Estimation of Cell Proliferation Dynamics Using CFSE Data

H.T. Banks,

Center for Research in Scientific Computation and Center for Quantitative Sciences in Biomedicine, North Carolina State University, Raleigh, NC 27695-8212

Karyn L. Sutton,

Center for Research in Scientific Computation and Center for Quantitative Sciences in Biomedicine, North Carolina State University, Raleigh, NC 27695-8212

W. Clayton Thompson,

Center for Research in Scientific Computation and Center for Quantitative Sciences in Biomedicine, North Carolina State University, Raleigh, NC 27695-8212

Gennady Bocharov,

Institute of Numerical Mathematics, RAS, Moscow, Russia

Marie Doumic,

INRIA Rocquencourt, Projet BANG, Domaine de Voluceau, 78153 Rocquencourt, France

Tim Schenkel,

Department of Virology, Saarland University, D-66421 Homburg, Germany

Jordi Argilaguet,

ICREA Infection Biology Lab, Dept of Experimental and Health Sciences, Univ. Pompeu Fabra, 08003 Barcelona, Spain

Sandra Giest,

ICREA Infection Biology Lab, Dept of Experimental and Health Sciences, Univ. Pompeu Fabra, 08003 Barcelona, Spain

Cristina Peligero, and

ICREA Infection Biology Lab, Dept of Experimental and Health Sciences, Univ. Pompeu Fabra, 08003 Barcelona, Spain

Andreas Meyerhans

Department of Virology, Saarland University, D-66421 Homburg, Germany

ICREA Infection Biology Lab, Dept of Experimental and Health Sciences, Univ. Pompeu Fabra, 08003 Barcelona, Spain

Abstract

CFSE analysis of a proliferating cell population is a popular tool for the study of cell division and division-linked changes in cell behavior. Recently [13, 43, 45], a partial differential equation (PDE) model to describe lymphocyte dynamics in a CFSE proliferation assay was proposed. We

© 2011 Elsevier B.V. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

present a significant revision of this model which improves the physiological understanding of several parameters. Namely, the parameter γ used previously as a heuristic explanation for the dilution of CFSE dye by cell division is replaced with a more physical component, cellular autofluorescence. The rate at which label decays is also quantified using a Gompertz decay process. We then demonstrate a revised method of fitting the model to the commonly used histogram representation of the data. It is shown that these improvements result in a model with a strong physiological basis which is fully capable of replicating the behavior observed in the data.

Keywords

Cell proliferation; cell division number; CFSE; label structured population dynamics; partial differential equations; inverse problems

1 Introduction

The quantitative analysis of cell division is an important problem at the intersection of biology and mathematics. Of the myriad applications and active areas of research, 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. Understanding cell proliferation is important 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. 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, as well as any division-linked changes which are observed. Thus the problem 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 cells. Second, there is a need for mathematical models which can describe the data obtained from such a procedure.

Unlike many other cell types, the inherent mobility in vivo and nonadherence in vitro of lymphocytes makes the accurate determination of lineage very challenging [47] (although new techniques have been developed for this purpose [34]). 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 [52, 63]. Early techniques, such as tritiated thymidine 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 [48]. 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 [52, 63].

Since it was first described in 1994 [48], serial dilution of the fluorescent dye carboxyfluorescein succinimidyl ester (CFSE) has become the de facto method for the determination of such cellular division histories. CFSE is nonradioactive and stably incorporated (so that measurable concentrations of CFSE 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) [48, 52, 54, 63]. Using a flow cytometer, the fluorescence intensity (a surrogate for CFSE content) of individual cells can

be measured. Because the CFSE content of a cell is divided approximately in half each time the cell divides, the number of divisions a cell has undergone can then be determined by comparing the measured amount of CFSE to the original CFSE content of an undivided cell [47, 48, 52, 54]. When individual cell fluorescent intensity measurements are binned into a histogram, each generation of cells appears as a “peak” in the histogram data.

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 [47, 48, 54, 63, 65]. Traditionally, a quantitative analysis of the population is possible if the peaks in the histogram data (see Figure 1) are fitted with gaussian or log-normal curves to determine the numbers of cells in each generation [47, 54]. While these basic frameworks provide an efficient measurement of the gross behavior of a proliferating culture, a more complete analysis (as well as comparisons among different cultures and extracellular conditions) requires a more extensive mathematical framework to establish a quantitative measure of the proliferation dynamics of the population.

Detailed mathematical analysis for asynchronously proliferating populations of cells began in earnest with the work of Gett and Hodgkin [24, 30]. By fitting CFSE histogram data with a series of log-normal curves, they computed cell numbers for each generation over the course of several days and tracked how these numbers changed over time. The authors go on to compute parameters such as mean division rate and mean time to first division. Because these models establish parameters which are readily identifiable from data sets, changes in parameters resulting from changing experimental conditions have been used to describe in exact terms how changing stimulatory and costimulatory conditions directly affect proliferative capacity.

Other models to describe cell population dynamics have focused, to varying degrees, on mathematical formalisms of cell cycle dynamics. Several authors have used linear compartmental ordinary differential equations (ODEs) [55, 62] to predict the number of cells in each generation as a function of time. More commonly, the Smith-Martin [58] model, in which the cell cycle is divided into a stochastic G1 state and a deterministic S-G2-M state, is mathematized as a system of delay differential equations [22, 23, 29, 41, 51]. Delay differential equations have the advantage of establishing a minimum cell cycle clock [22, 23, 29] to prevent the rapid onset of proliferation, although it is not clear if this is necessary [2].

There are alternatives to differential equation models of cell-cycle based cell proliferation dynamics. The cyton model, initially proposed by Hawkins et al., [32], assumes that a cell has a fixed time-to-divide and time-to-die, both chosen from fixed probability distributions; whichever parameter is smaller determines the fate of the cell in that generation, and with each division these two parameters are reset. Software implementing the method is freely available [33]. Another approach to the cell-cycle based modeling of a proliferating population is the use of a stochastic Bellman-Harris type branching process [36, 37, 59, 66], which is similar to the cyton model.

In general, all of these models are based upon estimation from the cell numbers computed by fitting 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 static distribution of cells in the population. However, the imposition of particular shapes for the generational structure of CFSE histogram data 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 development of CFSE and flow cytometry proliferation assays makes structured population models a natural framework in which to work. Significant literature exists on the subject of structured population models [18, 50, 57], and they have been widely applied to populations of cells [1, 16, 17, 26, 27, 53].

The measurement of CFSE fluorescence intensity (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 [19]. To our knowledge, Luzyanina, et al., [45], 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. 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. More recent work [6, 13, 43] has consistently demonstrated that this PDE framework can accurately model the observed histogram data from a CFSE-based proliferation assay. 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, a potential advantage when working with heterogeneous cell populations.

While this model is 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 [13]. In this document, we revisit the work presented in [13] and [45] primarily focusing on two key problems addressed there but not resolved. First, we address the issue with 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, we investigate the functional form for the rate at which label naturally decays. 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 [45] 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 and as such, might be used in a diagnostic setting [28] (e.g., to distinguish between healthy and diseased or abnormal cells based upon estimated proliferation rates), or to compare the effects of extracellular conditions such as stimulation strength and duration on proliferative behavior and cell viability [24, 30].

2 CFSE Data Set

For the analysis here we use the same data set as in [13, 45]. This data set is the result of an in vitro proliferation assay with human blood mononuclear cells (PBMCs) taken from a healthy blood donor. Briefly, approximately 5×10^6 to 5×10^7 were stained with $5\mu M$

CFDA-SE and stimulated to divide with 2.5 $\mu\text{g/mL}$ phytohaemagglutinin (PHA). The cells are placed in well plates at a density of 1×10^6 cells per milliliter of RPMI-1640/10% fetal calf serum (FCS) nutrient medium. Every 24 hours, cells from a single well are harvested and transferred to Trucount tubes containing 51466 beads. These cells are then stained with fluorescently labeled anti-CD4 antibodies and analyzed via flow cytometry. More detailed information on the experimental protocol used for this particular data set can be found in [13, 45]. More information regarding the protocol in general can be found in [47, 48, 54, 63, 65].

At each sample time, a fraction of the population of cells placed in the Trucount tube and stained with fluorescently labeled antibodies are counted by the cell sorter. This subpopulation contains all cell types present in a PBMC culture (CD4+ and CD8+ T cells, B cells, monocytes, etc.), however only CD4+ T cells are considered for mathematical analysis after gating them based on size, granularity and CD4+ expression. Because the entire contents of the tube are not collected, the cell counts obtained from the cytometer are scaled upward by the ratio of known beads in the tube to the number of beads actually counted.

Qualitatively, the flow cytometer returns a measure of the fluorescence intensity of a given cell, owing primarily to the presence of CFSE within the cell. It is known that measured CFSE FI has a linear relationship with the concentration of CFSE used in the staining process [47, Fig. 3] and is expected to correlate with the mass of CFSE within a cell. Because CFSE FI divides approximately in half with each subsequent division (at least for the first few generations; see Section 3) it is most convenient to use a logarithmic scale for CFSE FI. The most common representation for CFSE FI data is a series of histograms. (While it is possible, in general, to choose the bins for the histograms as one wishes, we remark that the current data set was already reported as histogram data when we began efforts on it.) The current data set is shown in Figure 1. It is this histogram data for which we seek a mathematical model. Each peak in the histogram data consists of cells which have undergone the same number of divisions. 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 [13] were generated by fitting to the data at all five points in time ($t = 24, 48, 72, 96$, 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 in culture cannot increase over the course of the experiment. This mass can decrease as a result of cell death and the natural decay/catabolism of CFSE within a cell. (While the separate measurements in time are obtained from distinct populations of cells in separate wells, the assumption that each well contains a sufficiently similar 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’ (the sum of all cells in the histogram, weighted by the measured FI) is greater at $t = 72$ hours than at the previous time point (see Figure 2), indicating a net increase in the mass of CFSE between $t = 48$ hours and $t = 72$ hours, a physical impossibility. The causes of this increase are not currently known. It is possible that this unusual result is explained by naturally occurring fluctuations in the number of beads counted by the cytometer, or possibly by the presence of additional cells types (e.g., monocytes) in the data. This unusual feature is present (at various measurement times) in several additional data sets which have been collected. As such, this feature will need to be addressed in the future by a more accurate observation model. At any rate, we assume for the current report that this anomaly is the result of measurement or scaling error or some

unknown and unmodeled biological event; hence the $t = 72$ hours data point will not be included in the present investigation. The new mathematical model proposed in this report, as well as the mathematical model from [13], have both been calibrated with and without the data point at $t = 72$ hours. The results (unpublished) confirm our suspicion as both models (which are derived from CFSE mass-conservation principles) are incapable of accounting for the erroneously large quantity of CFSE observed at $t = 72$ hours.

3 Mathematical Model

We begin this section by first recalling the previous PDE models [13, 45] proposed to describe the data. Let $n(t, z)$ be the label-structured population density indicating the density of cells at time t (hours) and log label intensity z (log units of intensity, log UI). In the analysis of [13], it was shown that the most effective mathematical model had the form

$$\frac{\partial n(t, z)}{\partial t} + \frac{\partial [cn(t, z)]}{\partial z} = -(\alpha(t, z+ct) + \beta(z+ct)) n(t, z) + \chi[z_{\min}, z_{\max} - \log \gamma] 2\gamma \alpha(t, z+ct + \log \gamma) n(t, z + \log \gamma), \quad (1)$$

where $\alpha(t, s)$ is the rate of cell proliferation (hr^{-1}), $\beta(s)$ is the rate of cell death (hr^{-1}), c is the rate at which label is lost (log UI hr^{-1}), and γ is the ratio of CFSE FI of a mother cell to that of a daughter cell. Thus the second term on the left represents velocity of decay of the fluorescence label while the last term on the right represents rate of production of new cells due to cell division. A complete derivation and detailed explanations are given in [6, 13].

It was shown in [13] that the model (1) is quite capable of providing an accurate fit to the data set at hand. However, the parameter γ was used to represent an unknown process responsible for determining the CFSE FI of two daughter cells given the CFSE FI of a mother cell. It was not known at the time what processes might be represented by γ or how that parameter should be interpreted (see also [45] where γ was first introduced). We would like to make this model more physically relevant by explaining the mechanism underlying the parameter γ .

Mathematically, 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, 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 [13] was $\gamma^* = 1.5169$, implying the creation of label at division. Similar results were also obtained in [45]. 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.

One possible solution conjectured in [13] 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 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 [48, 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 [47], 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 [48, 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 Section 2, 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, then the total fluorescence intensity of a cell is

$$X_i = X_i^{\text{CFSE}} + X^{\text{Auto}}. \quad (2)$$

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

$$X_{i+1} = X_i^{\text{CFSE}} / 2 + X^{\text{Auto}}. \quad (3)$$

Contrary to previous interpretations of the parameter γ , one can see from Equation (3) 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 $X_i^{\text{CFSE}} \gg X^{\text{Auto}}$, 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 5.

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 [35], and a formula corresponding to (3) above appears in [47]. However, autofluorescence has not been used in previous PDE formulations [13, 43, 45] to describe the dilution of CFSE by division. Thus the incorporation of AutoFI into the mathematical model is an important revision to the physiological basis of the PDE model.

Based upon this discussion, the model (1) can be replaced by the updated model (in the original intensity coordinates x)

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

As before, it is assumed that the population of cells can only change by division, death, or the nature decay of CFSE within a cell. Cells are assumed to partition CFSE evenly upon division without any instantaneous loss of label to the external environment. As before [13, 45], this model is derived from the mass-balance principles of the Bell-Anderson [18] and Sinko-Streifer [57] models. A complete list of assumptions as well as a derivation of the mathematical model can be found in the Technical Report version [14] of this paper.

Given this new model, 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. 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.

It was hypothesized in [45] that an exponential rate of loss is sufficient to model the label loss observed in the data. In order to validate 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.

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 [13, 45], we begin with a model of label loss that decays exponentially to the autofluorescence of unlabeled cells,

$$x_1(t) = (x(0) - x_a)e^{-ct} + x_a. \quad (5)$$

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 [39, pg. 12]

$$x_2(t) = (x(0) - x_a) \exp\left(-\frac{c}{k}(1 - e^{-kt})\right) + x_a. \quad (6)$$

The loss rate function, of vital importance to the PDE formulation (4), is $v(t, x) = \frac{dx}{dt}$. Thus the equations (5 – 6) correspond to the loss rate functions

$$v_1(x) = -c(x - x_a), \quad (7)$$

and

$$v_2(t, x) = -c(x - x_a)e^{-kt}, \quad (8)$$

respectively. We remark that (6) is a generalization of (5), the latter being the limiting value (as $k \rightarrow 0$) 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 collected from nonproliferating cells. On the other hand, the parameter x_a appears in other parts of the model (4), and it seems possible to conclude that this parameter would be readily identifiable once incorporated into (4) with the full proliferation 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 [46, Figure 1], [48, Figure 2] and [54, 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 are depicted in Figure 3. It is clear from the figure that the Gompertz model provides a more accurate description of the rate of CFSE decay observed in the experimental data. More details regarding these two data sets as well as comments and detailed results for the least-squares procedure are contained in the Technical Report version [14] of this paper.

We 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 (5) that the exponential model predicts $x \rightarrow x_a$ as $t \rightarrow \infty$. However, it is known that CFSE stained cells retain detectable fluorescence for up to several weeks in vivo [48]. 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. 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 [52]. These processes combine to produce the commonly observed “biphasic decay” of CFSE FI over time [49, 63]. 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 [39].

4 Parameter Estimation Procedure

With the primary features of the revised model now addressed, we are ready to validate the model with data. The current model, accounting for CFSE AutoFI and the Gompertz decay of the label, is

$$\frac{\partial n(t, x)}{\partial t} - ce^{-kt} \frac{\partial [(x - x_a) n(t, x)]}{\partial x} = -(\alpha(t, x) + \beta(t, x)) n(t, x) + \chi_{[x_a, x^*]} 4\alpha(t, 2x - x_a) n(t, 2x - x_a). \quad (9)$$

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

$$\begin{aligned} n(t, x_{\max}) &= 0 \\ v(t, x_a) n(t, x_a) &= 0. \end{aligned} \quad (10)$$

Note from (8) we have $v(t, x_a) = 0$ for all t , and hence the left boundary condition is trivially satisfied. Finally, we assume we are given some initial condition

$$n(0, x) = \Phi(x), \quad (11)$$

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

4.1 Computational Considerations

While the model (9) 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 as a result of a highly restrictive CFL condition for stability. Thus, 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 [13, 45], it is less effective here because of the different form of the label loss rate function. 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 (9) and is needed to conserve the total label in the population. With this change of variables the new PDE model is

$$\frac{\partial \tilde{n}}{\partial t} - c e^{-kt} \frac{\partial}{\partial y} \left[\frac{\tilde{n}(t, y)}{\ln 10} \right] = -(\tilde{\alpha}(t, y) + \tilde{\beta}(t, y) - c e^{-kt}) \tilde{n}(t, y) + \chi_{(-\infty, y^*]} 2\tilde{\alpha}(t, y + \log_{10} 2) \tilde{n}(t, y + \log_{10} 2), \quad (12)$$

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 $\Phi(y) = 10^y \ln(10) \Phi(t, 10^y + x_a)$. The right boundary condition follows immediately from (10) while the left boundary has been removed to $y = -\infty$. We remark that the CFL condition for the transformed model (12) is significantly easier to satisfy so that the computations proceed much faster. While (12) 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 (12) 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 (12) on the same computational domain regardless of the parameters (i.e., those passed in by the nonlinear optimization solver). For the current data set, the forward solution is computed using a publicly available hyperbolic PDE solver written by L. Shampine which implements the Lax-Wendroff finite-difference scheme. We use 512 evenly spaced nodes in the interval $y \in [0, 3.5]$.

It should be noted that 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 (12) 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. The primary issues involve the determination of the initial condition from the data, and the comparison of the computed model solution to the data.

As discussed in Section 2, the cytometry data is reported as a time-series of histograms showing the numbers of cells counted into a given bin corresponding to a particular range of log CFSE FI values. That is, the histogram measures cells in the $z = \log_{10} x$ coordinate. While it is, in general, possible for the experimenter to set the bins to his/her own liking, we again recall that the current data set was obtained with bins already set. The original data set, as used in [13], was taken at 24 hour intervals over the course of 6 days. At time t_i , the data is stored as a set of ordered pairs (z_i^j, n_i^j) , $j=1, \dots, J(i)$ (the notation is meant to emphasize that the bins change each day). This ordered pair corresponds to the number of cells n_i^j counted into the bin with left boundary z_i^j . (As a consequence, notice that we are unable to determine the width of the right-most bin at each time point; these points are simply removed from the data set).

Data from Day 0 ($t = 0$ hours) is used to form the initial condition for the model. First, we drew a smooth line through the data; ordered pairs representing the line were then determined using DataThief [60]. These smoothed histogram curves were then scaled upward into a smooth initial condition density $\Phi(z)$ so that the total label content (mass of CFSE) is the same for the smooth density as for the original histogram data. Finally, given the initial condition $\Phi(z)$, we transform this into an initial condition for $\tilde{n}(t, y)$ by noting that $y = \log_{10}(10^z - x_a)$ and using the label-preserving identity

$$\widehat{\Phi}(z) = \frac{10^z}{10^z - x_a} \tilde{\Phi}(y(z)). \quad (13)$$

This, then, provides an initial condition $\tilde{\Phi}(y)$ for (12).

Data from the remaining 5 days is used to fit the model to the data. In order to compare the model, a density defined in terms of y , to the data, a histogram in terms of z , we need to perform two steps. First, we must transform the structured density $\tilde{n}(t, y)$ (the solution to the model) into a structured density $\hat{n}(t, z)$, a function of z . This is done analogously to the transformation (13) above. Second, we must transform this structured density into histogram numbers. To do this, we note that at time t_i , the total number of cells with FI between z_i^j and z_i^{j+1} is

$$I[\hat{n}](t_i, z_j) \equiv \int_{z_i^j}^{z_i^{j+1}} \hat{n}(t_i, z) dz \approx \left[\frac{\hat{n}(t_i, z_i^{j+1}) + \hat{n}(t_i, z_i^j)}{2} \right] (z_i^{j+1} - z_i^j), \quad (14)$$

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. 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.

4.2 Parameterizations of α and β

We next turn our attention to the parameterizations of the functions $\alpha(t, y)$ and $\beta(t, y)$. 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 [11, 12]. Previous work has established that division-linked changes in proliferation and death rates are an important aspect of an accurate mathematical model [40, 44]. 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 [13, 43, 45]. 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 how the correlation between division number and structure variable is considered. As discussed at greater length in [6, 13], 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. Mathematically, it is shown in [6] that this is 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). Similar situations frequently arise in mechanics and fluid dynamics, where discussions of Eulerian and Lagrangian formulations abound. 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. For a cell with state y at time t , this is equivalent to finding the intersection of the y -axis with the characteristic line passing through (t, y) . Given the characteristic lines

$$\frac{dy}{dt} = \frac{-ce^{-kt}}{\ln 10},$$

it follows that the cell located at (t, y) was originally located (in the absence of division) at

$$s(t, y) = y + \frac{c}{k \ln 10} (1 - e^{-kt}).$$

It is shown in [6, 13] that the quantity $s(t, y)$ is more strongly correlated with division number than the quantity y . Moreover, the use of this 'translated coordinate' for the parameterization of α and β provided a more accurate model of the observed data when compared to the simple implementation of spatial dependence. It was hypothesized that the improvement resulted from a more direct association between division number and proliferation rate. However, that analysis was done with a different label loss function, and we desire to repeat the analysis of [13] with the new model (12) and new label loss function (equivalent to (8)). Thus we consider four different parameterizations of the proliferation rate α . In Section 5, 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

$$\alpha = \alpha(y) = \sum_{k=1}^{K_\alpha} a_k l_k^{(\alpha)}(y), \quad (15)$$

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.

We also consider the possibility that the proliferation rate depends explicitly on time. Indeed, we see in the data in Figure 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

$$\alpha = \alpha(t, y) = \sum_{k=1}^{K_\alpha} \sum_{m=1}^M a_{km} l_k^{(\alpha)}(y) l_m^{(t)}(t), \quad (16)$$

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 smoothness in the forward simulation 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 also 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

$$\alpha = \alpha(t, s) = \alpha(t, s(t, y)) = \sum_{k=1}^{K_\alpha} \sum_{m=1}^M a_{km} l_k^{(\alpha)}(s) l_m^{(t)}(t), \quad (17)$$

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

Results from [13, 43, 45] 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

$$\beta = \beta(y) = \sum_{k=1}^{K_\beta} b_k l_k^{(\beta)}(y). \quad (18)$$

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

$$\beta = \beta(s) = \beta(s(t, y)) = \sum_{k=1}^{K_\beta} b_k l_k^{(\beta)}(s) \quad (19)$$

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

4.3 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 $\hat{n}(t, y; \theta)$ satisfying (12) 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) [15, 20, 21], we assume that the data $\hat{n}_i^j$ 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 realizations

$$n_j^i = I[\hat{n}](t_i, z_j; \theta_0) + \varepsilon_{ij} \quad (20)$$

of a random variable $N_i^j = I[\hat{n}](t_i, z_j; \theta_0) + \varepsilon_{ij}$. It is assumed that the random variables $\{\varepsilon_{ij}\}$, representing noise and/or measurement error, satisfy $E[\varepsilon_{ij}] = 0$ and $Var(\varepsilon_{ij}) = \sigma^2$. 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 [7, 15, 56]. Such a more detailed analysis will be necessary in order to establish meaningful confidence intervals for estimated parameters in future work [7]. 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.

The goal of the OLS procedure is the determination of the parameter θ which minimizes the sum of squared residuals. Thus we seek to determine the estimate

$$\hat{\theta}_{OLS} = \arg \min_{\theta \in \Theta} \sum_{i=1}^I \sum_{j=1}^{J(i)} (I[\hat{n}](t_i, z_j; \theta) - n_j^i)^2 = \arg \min J(\theta), \quad (21)$$

where Θ is a set of admissible parameters for the model (see Table 1). 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. More information regarding the statistical aspects of this inverse problem, with implications for asymptotic analysis and standard error/confidence interval estimation, can be found in the Technical Report version [14] of this paper.

5 Results

The primary uncertainty in the inverse problem procedure is the choice of nodes $\{y_k\}$ for the estimation of the proliferation rate. (The death rate is not expected to vary as significantly with division number.) We have no a priori information as to how many nodes should be used nor any information regarding where those nodes should be placed. After some trial and error, we have found 13 evenly spaced nodes in the interval $[1.125, 2.925]$ to provide the best results. This seems to be an appropriate balance between a decrease in OLS cost (as the number of parameters is increased) and the desire to estimate a somewhat regular proliferation rate function (a so-called ‘regularization by discretization’ [10, 11]; see Figure 8 in the Appendix). On the surface, it appears to be a potential drawback of our method that the estimated proliferation rate may change as the number and placement of nodes used for the estimation changes. However, the proliferation rate function itself is not of immediate use to biologists. Rather, meaningful proliferation rate estimates should indicate the average rate of proliferation in terms of the numbers of divisions undergone. After using Figure 1 to determine approximate ranges corresponding to particular generations of cells, the average value of the proliferation rate function can be determined in each range. These generation-specific rates are consistently estimated regardless of the choice of nodes (within reason) for the structural discretization of the proliferation rate (Table 2 in the Appendix).

We begin by calibrating the model when the proliferation rate is assumed to be independent of time. The OLS best-fit solution obtained using 13 nodes for the estimation of the proliferation rate $\alpha(y)$ and 5 nodes for the estimation of the death rate $\beta(y)$ is shown in Figure 4. More detailed results, including a complete list of estimated parameter values, can be located in the Technical Report version [14] of this paper. The total OLS cost for the estimation is $J(\hat{\theta}_{OLS}) = 1.7270 \times 10^{12}$.

We observe that, while the OLS best-fit solution is accurate for $t = 120$ 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 addition to the arguments of Section 4.2, 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 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 [13, 45] 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 4.2) in order to better estimate the numbers of cells in each generation at a given time. We would like to make use of residual-sum-of-squares based statistical tests for model refinements [7, 8, 15] 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 $J(\hat{\theta}_{OLS}) = 3.1302 \times 10^{11}$. The fit of the model to the data is shown in Figure 9 in the appendix, along with the estimated proliferation and death rate functions. Additional details regarding the estimated parameters can be found in the Technical Report [14]. There is a significant improvement from using the time-dependent proliferation rate when fitting the model to the data. 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 [15, 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, 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. Next, following the analysis of [13] and the discussion of Section 4.2, 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 estimated functions are analyzed in terms of division number (in the manner of Table 2 in the Appendix). It is worth noting 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 the model comparison tests described above are not directly 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 addition to providing a slightly lower cost!), which justifies its use.

As before, 73 parameters arise when the model is parameterized in terms of the translated variable s . After calibration, the OLS cost of this parameterization is $J(\hat{\theta}_{OLS}) = 3.0901 \times 10^{11}$, and the fit of the model to the data is shown in Figure 5. The estimated AutoFI parameters is $x_a = 6.4053$, and the estimated Gompertz label loss parameters are $c = 5.5246 \times 10^{-3}$ and $k = 5.0323 \times 10^{-4}$. The estimated proliferation and death rate functions can be found Figure 10 in the Appendix, with exact numerical values for the nodes and the estimated numerical values of the rates at those nodes contained in the Technical Report [14].

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 are shown in the translated coordinate in Figure 6. While there is still some overlap among the generations of cells in the histogram data, the translated coordinate provides an axis on which cells do not drift as they slowly lose CFSE FI. Particularly when compared to Figure 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 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 7, thus preserving what we believe to be an important feature of the population of cells (that the proliferation rates change in time).

6 Discussion

In this document, we have presented significant modifications and clarifications to the results of [13] and [45]. Of primary importance is that the parameter γ , used in the previous model 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 the same data as in [13]. 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.

We also revise the manner in which the model, a structured density, is compared to the data, a series of histograms. Because the data used in this report is already in histogram form, the irregular size of certain bins caused some computational difficulty. In the future experiments, direct control over the histogram bin spacing will remove these difficulties. We also hope to use future data sets to determine an accurate statistical model for the error/noise in a given data set, and how that model changes as the histogram bins are selected in different ways. As the ultimate goal of any model of the immune response is the comparison of changing intra- and extracellular conditions on proliferative behavior [30, 31], 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 [7, 8, 20, 56] exist, but rely upon a correct underlying statistical model assumption for the data. While we assume a constant variance model here (20), it has been shown that this is not correct [13], and thus any confidence intervals computed from such a formulation would be in error. (This is not to say that the estimated parameters reported here are invalid, but only that we cannot continue further to quantify the certainty of those estimates without additional information.) 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 [15].

Similarly, the mechanism responsible for the apparently spurious measurement at $t = 72$ hours must be properly accounted for in future work. As has been shown, the total mass of CFSE in the cell culture is measured to increase from $t = 48$ hours to $t = 72$ hours in [13]. 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 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 been noticed. Thus, in addition to the statistical model discussed in the previous paragraph, future work will need to focus on establishing an observation model to accurately account for the manner in which the cell population data is represented.

In support of the results of [13], 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 6), the peaks

corresponding to distinct division numbers align much more closely than when presented in terms of the measurement variable z (Figure 1). As such, the estimation of the proliferation and death rate functions in terms of this variable should more directly relate division number to the rate at which cells in a given generation divide. It follows that average rates of proliferation can be calculated in terms of the number of divisions a cell has undergone, and the dependence of these rates on time can also be explored (Table 7). Thus, we find that the translated coordinate s , as a result of its strong correlation with division number, 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.

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, this best-fit results presented in this report require 73 parameters. Optimization times range from 1 to 8 hours, depending upon the accuracy of the initial iterate and the 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, 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. For instance, we might generalize the model presented here to account for AutoFI which changes as a function of time (as cells are activated and/or enter a quiescent state) or which varies from cell to cell in the population. While the exact shape of the estimated proliferation and death rate functions may change with various choices of nodes (Figure 8), we have seen that, for reasonable parameterization, the average proliferation rates (as a function of division number) are consistently estimated (Table 2). 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 $\alpha(t, s)$.

In the current model formulation, the rates of proliferation and/or death are essentially

exponential (in the sense that the rate of change of the population, $\frac{\partial n}{\partial t}$ is directly proportional to the total population $n(t, y)$) with rates α and β , respectively. We do not need to make use of delays or minimum cell cycle times in order to fit the data (as the time-dependence of the function α accomplishes the same effect). An interesting generalization of the current model would be the incorporation of volume structure. As cells would need to progress from size V to $2V$ before division, this would naturally require the incorporation of some cell cycle time. Moreover, forward scatter (FSC) of laser light may possibly be used as an observable surrogate for cell size. Importantly, the current model assumes that all cells in the population behave in exactly the same manner. However, recent work has demonstrated that there are cohorts of closely related cells whose behavior is correlated in some way [25, 34, 61, 64]. Thus it may be necessary to examine in the framework of [4, 5, 9] the effects of probabilistically distributed parameters within the population. This same framework could be used to describe subpopulations of cells with varying levels of AutoFI (as discussed in the previous paragraph) or which internally process/catabolize intracellular die at different rates.

Using the data set from [13, 45], 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. 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 and various types of cells are subjected to a variety of experimental conditions and then measured. 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. 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 [28], and perhaps in a more integrated model of cell dynamics [38, 42].

Acknowledgments

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, in part by the Russian Foundation for Basic Research (Grant 11-01-00117), in part by the Program of the Russian Academy of Sciences Basic Research for Medicine, in part by the Agence Nationale de la Recherche, Grant No. ANR-09-BLAN-0218 TOPPAZ, in part by the Deutsche Forschungsgemeinschaft and in part by grant SAF2010-21336 from the Spanish Ministry of Science and Innovation. The authors are also grateful to referees for constructive comments and suggestions for improvements in this manuscript.

References

1. Arino O, Sanchez E, Webb GF. Necessary and sufficient conditions for asynchronous exponential growth in age structured cell populations with quiescence. *J. Mathematical Analysis and Applications*. 1977; 215:499–513.
2. Asquith B, Debacq C, Florins A, Gillet N, Sanchez-Alcaraz T, Mosley A, Willems L. Quantifying lymphocyte kinetics in vivo using carboxyfluorein diacetate succinimidyl ester. *Proc. R. Soc. B*. 2006; 273:1165–1171.
3. Aubin JE. Autofluorescence of viable cultured mammalian cells. *J. Histochem. Cytochem.* 1979; 27:36–43. [PubMed: 220325]
4. Banks, HT.; Botsford, LW.; Kappel, F.; Wang, C. *Proc. 2nd Course on Math. Ecology*. Singapore: World Scientific Press; 1988. Modeling and estimation in size structured population models, LCDS/CSS Report 87-13, Brown University March, 1987; p. 521-541.(Trieste, December 8–12, 1986)
5. Banks HT, Davis JL. A comparison of approximation methods for the estimation of probability distributions on parameters. *Appl. Num. Math.* 2007; 57:753–777.
6. Banks HT, Charles Frederique, Doumic Marie, Sutton Karyn L, Thompson W. Clayton. Label structured cell proliferation models. *Appl. Math. Letters*. 2010; 23:1412–1415.
7. Banks, HT.; Davidian, M.; Samuels, J.; Sutton, KL. An inverse problem statistical methodology summary, CRSC-TR08-01, NCSU, January, 2008. In: Chowell, G., et al., editors. *Mathematical and Statistical Estimation Approaches in Epidemiology*. New York: Berlin Heidelberg; 2009. Chapter 11. p. 249-302.
8. Banks, HT.; Fitzpatrick, BG. 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.; 1989. p. 262-273.
9. Banks HT, Fitzpatrick BF. Estimation of growth rate distributions in size-structured population models, CAMS Tech. Rep. 90-2, Univ. of Southern California, January, 1990. *Quart. Appl. Math.* 1991; 49:215–235.
10. Banks, HT.; Iles, DW. 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 PDE's; Springer Lecture Notes in Control and Inf. Sci.; Gainesville, Florida. 1987. p. 130-142.
11. Banks, HT.; Kunisch, K. Estimation Techniques for Distributed Parameter Systems. Boston: Birkhauser; 1989.
 12. Banks HT, Pedersen M. Well-posedness of inverse problems for systems with time dependent parameters, CRSC-TR08-10, NCSU, August 2008. Arab. J. Sci. Eng. Math. 2009; 1:39–58.
 13. Banks HT, Sutton KarynL, Thompson W. Clayton, Bocharov Gennady, Roose Dirk, Schenkel Tim, Meyerhans Andreas. Estimation of cell proliferation dynamics using CFSE data, CRSC-TR09-17, NCSU, August, 2009. Bull. Math. Biol. 2011; 70:116–150. [PubMed: 20195910]
 14. Banks HT, Sutton KL, Thompson WC, Bocharov G, Doumic M, Schenkel T, Argilaguet J, Giest S, Peligero C, Meyerhans A. A new model for the estimation of cell proliferation dynamics using CFSE data, CRSC-TR11-05, NCSU. 2011 July. (Revised).
 15. Banks, HT.; Tran, HT. Mathematical and Experimental Modeling of Physical and Biological Processes. Boca Raton London New York: CRC Press; 2009.
 16. Basse B, Baguley B, Marshall E, Wake G, Wall D. Modelling the flow cytometric data obtained from unperturbed human tumour cell lines: Parameter fitting and comparison. Bull. Math. Biol. 2005; 67:815–830. [PubMed: 15893554]
 17. Bekkal Brikci F, Clairambault J, Ribba B, Perthame B. An age-and-cyclin-structured cell population model for healthy and tumoral tissues. J. Math. Biol. 2008; 57:91–110. [PubMed: 18064465]
 18. Bell G, Anderson E. Cell Growth and Division I. A Mathematical Model with Applications to Cell Volume Distributions in Mammalian Suspension Cultures. Biophysical Journal. 1967; 7:329–351. [PubMed: 6069910]
 19. Bonhoeffer S, Mohri H, Ho D, Perelson AS. Quantification of cell turnover kinetics using 5-Bromo- 2'-deoxyuridine. J. Immunology. 2000; 64:5049–5054. [PubMed: 10799860]
 20. Carroll, RJ.; Ruppert, D. Transformation and Weighting in Regression. London: Chapman Hall; 2000.
 21. Davidian, M.; Giltinan, DM. Nonlinear Models for Repeated Measurement Data. London: Chapman and Hall; 2000.
 22. DeBoer RJ, Ganusov VV, Milutinovic D, Hodgkin PD, Perelson AS. Estimating lymphocyte division and death rates from CFSE data. Bull. Math. Biol. 2006; 68:1011–1031. [PubMed: 16832737]
 23. DeBoer RJ, Perelson Alan S. Estimating division and death rates from CFSE data. J. Comp. and Appl. Mathematics. 2005; 184:140–164.
 24. Deenick EK, Gett AV, Hodgkin PD. Stochastic model of T cell proliferation: a calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival. J. Immunology. 2003; 170:4963–4972. [PubMed: 12734339]
 25. Duffy K, Subramanian V. On the impact of correlation between collaterally consanguineous cells on lymphocyte population dynamics. J. Math. Biol. 2009; 59:255–285. [PubMed: 18956191]
 26. Farkas JZ. Stability conditions for the non-linear McKendrick equations. Appl. Math. and Comp. 2004; 156:771–777.
 27. Farkas JZ. Stability conditions for a non-linear size-structured model. Nonlinear Analysis: Real World Applications. 2005; 6:962–969.
 28. Fulcher DA, Wong SWJ. Carboxyfluorescein diacetate succinimidyl ester-based assays for assessment of T cell function in the diagnostic laboratory. Immunology and Cell Biology. 1999; 77:559–564. [PubMed: 10571678]
 29. Ganusov VV, Pilyugin SS, De Boer RJ, Murali-Krishna K, Ahmed R, Antia R. Quantifying cell turnover using CFSE data. J. Immunological Methods. 2005; 298:183–200.
 30. Gett AV, Hodgkin PD. A cellular calculus for signal integration by T cells. Nature Immunology. 2000; 1:239–244. [PubMed: 10973282]
 31. Hasbold J, Gett AV, Rush JS, Deenick E, Avery D, Jun J, Hodgkin PD. Quantitative analysis of lymphocyte proliferation and differentiation in vitro using carboxyfluorescein diacetate succinimidyl ester. Immunology and Cell Biology. 1999; 77:516–522. [PubMed: 10571672]

32. Hawkins ED, Hommel Mirja, Turner ML, Battye Francis, Markham J, Hodgkin PD. Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. *Nature Protocols*. 2007; 2:2057–2067.
33. Hawkins ED, Turner ML, Dowling MR, van Gend C, Hodgkin PD. A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proc. Natl. Acad. Sci.* 2007; 104:5032–5037. [PubMed: 17360353]
34. Hawkins ED, Markham JF, McGuinness LP, Hodgkin PD. A single-cell pedigree analysis of alternative stochastic lymphocyte fates. *Proc. Natl. Acad. Sci.* 2009; 106:13457–13462. [PubMed: 19633185]
35. Hodgkin PD, Lee J, Lyons AB. B cell differentiation and isotype switching is related to division cycle number. *J. Exp. Med.* 1996; 184:277–281. [PubMed: 8691143]
36. Hyrien O, Zand MS. A mixture model with dependent observations for the analysis of CFSE-labeling experiments. *J. American Statistical Association*. 2008; 103:222–239.
37. Hyrien O, Chen R, Zand MS. An age-dependent branching process model for the analysis of CFSE-labeling experiments. *Biology Direct*. 2010; 2 Published Online.
38. Kirschner DE, Chang ST, Riggs TW, Perry N, Linderman JJ. Toward a multiscale model of antigen presentation in immunity. *Immunological Reviews*. 2007; 216:93–118. [PubMed: 17367337]
39. Kot, M. *Elements of Mathematical Ecology*. Cambridge UP: Cambridge, UK: 2001.
40. Lee HY, Perelson AS. Modeling T cell proliferation and death in vitro based on labeling data: generalizations of the Smith-Martin cell cycle model. *Bull. Math. Biol.* 2008; 70:21–44. [PubMed: 17701260]
41. Leon K, Faro J, Carneiro J. A general mathematical framework to model generation structure in a population of asynchronously dividing cells. *J. Theoretical Biology*. 2004; 229:455–476.
42. Louzoun Y. The evolution of mathematical immunology. *Immunological Reviews*. 2007; 216:9–20. [PubMed: 17367331]
43. Luzyanina T, Roose D, Bocharov G. Distributed parameter identification for a label-structured cell population dynamics model using CFSE histogram time-series data. *J. Math. Biol.* 2009; 59:581–603. [PubMed: 19096849]
44. Luzyanina T, Mrusek M, Edwards JT, Roose D, Ehl S, Bocharov G. Computational analysis of CFSE proliferation assay. *J. Math. Biol.* 2007; 54:57–89. [PubMed: 17093999]
45. Luzyanina T, Roose D, Schenkel T, Sester M, Ehl S, Meyerhans A, Bocharov G. Numerical modelling of label-structured cell population growth using CFSE distribution data. *Theoretical Biology and Medical Modelling*. 2007; 4 Published Online.
46. Lyons AB. Analysing cell division in vivo and in vitro using flow cytometric measurement of CFSE dye dilution. *J. Immunological Methods*. 2000; 243:147–154.
47. Lyons AB, Hasbold J, Hodgkin PD. Flow cytometric analysis of cell division history using dilution of carboxyfluorescein diacetate succinimidyl ester, a stably integrated fluorescent probe. *Methods in Cell Biology*. 2001; 63:375–398. [PubMed: 11060850]
48. Lyons AB, Parish CR. Determination of lymphocyte division by flow cytometry. *J. Immunol. Methods*. 1994; 171:131–137. [PubMed: 8176234]
49. Matera G, Lupi M, Ubezio P. Heterogeneous cell response to topotecan in a CFSE-based proliferative test. *Cytometry A*. 2004; 62:118–128. [PubMed: 15536634]
50. Metz JA, Diekmann O. The dynamics of physiologically structured populations. *Lecture Notes in Biomathematics*. 1986; 68
51. Nordon RE, Nakamura M, Ramirez C, Odell R. Analysis of growth kinetics by division tracking. *Immunology and Cell Biology*. 1999; 77:523–529. [PubMed: 10571673]
52. Parish C. Fluorescent dyes for lymphocyte migration and proliferation studies. *Immunology and Cell Biol.* 1999; 77:499–508. [PubMed: 10571670]
53. Perthame, B. *Transport Equations in Biology*. Birkhauser Frontiers in Mathematics, Basel; 2007.
54. Quah B, Warren H, Parish C. Monitoring lymphocyte proliferation in vitro and in vivo with the intracellular fluorescent dye carboxyfluorescein diacetate succinimidyl ester. *Nature Protocols*. 2007; 2(9):2049–2056.

55. Revy P, Sospedra M, Barbour B, Trautmann A. Functional antigen-independent synapses formed between T cells and dendritic cells. *Nature Immunology*. 2001; 2:925–931. [PubMed: 11561183]
56. Sever, GA.; Wild, CJ. *Nonlinear Regression*. Hoboken: Wiley; 2003.
57. Sinko J, Streifer W. A New Model for Age-Size Structure of a Population. *Ecology*. 1967; 48:910–918.
58. Smith JA, Martin L. Do Cells Cycle? *Proc. Natl. Acad. Sci.* 1973; 70:1263–1267. [PubMed: 4515625]
59. Subramanian VG, Duffy KR, Turner ML, Hodgkin PD. Determining the expected variability of immune responses using the cyton model. *J. Math. Biol.* 2008; 56:861–892. [PubMed: 17982747]
60. Tummers B, Thief Data III. 2006 <http://datathief.org/>.
61. Turner ML, Hawkins ED, Hodgkin PD. Quantitative regulation of B cell division destiny by signal strength. *J. Immunology*. 2009; 181:374–382. [PubMed: 18566403]
62. Veiga-Fernandez H, Walter U, Bourgeois C, McLean A, Rocha B. Response of naive and memory CD8+ T cells to antigen stimulation in vivo. *Nature Immunology*. 2000; 1:47–53. [PubMed: 10881174]
63. Wallace PK, Tario JD Jr, Fisher JL, Wallace SS, Ernstoff MS, Muirhead KA. Tracking antigen-driven responses by flow cytometry: monitoring proliferation by dye dilution. *Cytometry A*. 2008; 73:1019–1034. [PubMed: 18785636]
64. Wellard C, Markham J, Hawkins ED, Hodgkin PD. The effect of correlations on the population dynamics of lymphocytes. *J. Theoretical Biology*. 2010; 264:443–449.
65. Witkowski JM. Advanced application of CFSE for cellular tracking. *Current Protocols in Cytometry*. 2008;9.25.1–9.25.8. [PubMed: 18770647]
66. Yates A, Chan C, Strid J, Moon S, Callard R, George AJT, Stark J. Reconstruction of cell population dynamics using CFSE. *BMC Bioinformatics*. 2007; 8 Published Online.

Figure 1.
Original histogram data from [13, 45].

Figure 2. Total label content $\int zn(t, z)dz$ over time for the data from [13]. The increase at $t = 72$ hours is a physiological impossibility.

Figure 3. Results of fitting the exponential model (5) and the Gompertz model (6) to the mean CFSE data. For both Donor 1 (top) and Donor 2 (bottom), we see that the Gompertz model is more capable of accurately replicating the observed data.

Figure 4.

OLS best-fit solution with $\alpha = \alpha(y)$ (13 nodes), $\beta = \beta(y)$ (5 nodes). 21 total parameters in the model, total cost $J(\hat{\theta}_{OLS}) = 1.7270 \times 10^{12}$. 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.

Figure 5.

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

Figure 6.

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.

Figure 7.
Average proliferation rate as a function of time for each generation of cells.

Figure 8.

Left: Estimated proliferation rate function for three different 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). It is these computed average rates of proliferation that are of biological interest, and these rates are consistently estimated regardless of the parameterization used. Right: the corresponding estimated death rate function $\beta(y)$, estimated using 5 fixed nodes in each case.

Figure 9.

OLS best-fit solution with $\alpha = \alpha(t, y)$, $\beta = \beta(y)$. 73 total parameters in the model, total cost $J(\hat{\theta}_{OLS}) = 3.1302 \times 10^{11}$. The use of a time-dependent proliferation rate results in a significant improvement in the fit of the model to the data.

Figure 10.

OLS best-fit proliferation and death rate function $\alpha(s, t)$ (top) and $\beta(s)$ (bottom), respectively, when the proliferation rate is assumed to depend on both s and t . The parameterization in terms of s , a coordinate which correlates very strongly with division number, reduces the possibility of error or bias when average proliferation rates (in terms of division number) are computed (Figure 7).

Table 1

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

Table 2

Average proliferation rates (in units 1/hr) in terms of numbers of divisions undergone, computed from Figure 8. Using Figure 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