

Mathematical and Statistical Model Misspecifications in Modeling Immune Response in Renal Transplant Recipients

H.T. Banks, R.A. Everett, Shuhua Hu, Neha Murad, and H.T. Tran
Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC 27695-8212 USA

November 28, 2016

Abstract

We examine uncertainty in clinical data from a kidney transplant recipient infected with BK virus and investigate *mathematical model and statistical model misspecifications* in the context of least squares methodology. A difference-based method is directly applied to data to determine the correct statistical model that represents the uncertainty in data. We then carry out an inverse problem with the corresponding generalized least squares technique and use the resulting residual plots to detect mathematical model discrepancy. This process is implemented using both clinical and simulated data. Our results demonstrate mathematical model misspecification when both simpler and more complex models are assumed compared to data dynamics.

Key words: renal transplant, BK virus, polyomavirus nephropathy, inverse problems, pseudo measurement errors, difference-based methods, model misspecification, statistical error model

1 Introduction

Kidneys are an important pair of organs that extract waste from the blood, regulate body fluids, form urine, and aid in other important bodily functions. Blood flows into tiny blood vessel clusters in the kidney, called glomeruli, where the waste is filtered out to become urine. Glomerular Filtration Rate (GFR) is often used as an indicator for kidney health and function; it measures the rate at which the kidney clears toxic waste from the blood. A GFR number of 90 or less in adults is used as an indicator for kidney disease [12]. Chronic kidney disease (CKD), also commonly known as chronic kidney failure, is characterized by gradual but progressive loss of kidney function. The fifth stage of CKD, called End Stage Renal Disease (ESRD), occurs when kidney function reduces to less than 15% and leads to permanent kidney failure [12]. Patients with ESRD have two choices of therapy - dialysis or kidney transplantation. Kidney transplantation is often chosen since transplants (grafts) can improve survival and lower healthcare costs compared to dialysis [16]. As of November 2016, there are currently 121,678 people waiting for lifesaving organ transplants in the U.S., of which 100,791 await kidney transplants [12]. Donor kidneys can originate from either living or deceased donors. In 2011- 2012, 50.8% of patients who received deceased donor transplants experienced graft failure at 10 years, compared to 34.7% for those receiving a living donor transplant [13].

Kidney rejection is one of the most common causes for renal graft failures (64%), followed by infections such as polyomavirus-associated nephropathy (PVAN) (7%) [8]. In order to prevent the body from rejecting the transplant, patients are generally required to take immunosuppressants for their remaining life. However this usually leaves the patient susceptible to various bacterial and viral pathogens and can even reactivate latent viruses preexisting in the recipient and/or donor's organ. Common viruses that impact transplant recipients include human cytomegalovirus (HCMV), Epstein-Barr virus (EBV), human herpes virus (HHV)-6, HHV-7, and human polyomavirus 1 (BK virus) [3]. Thus for a renal transplant to be successful, a crucial but fragile balance needs to be struck between over-suppression and under-suppression of the immune system. While the former can weaken the body's immune response making it susceptible to infections, the latter can cause the immune response to fight the renal graft leading to kidney rejection.

Mathematical and statistical models are useful tools to investigate the cellular mechanisms of this complex biological process. Funk *et al.* [9] calculate the BK virus (BKV) clearance rate, virus half life, and loss of infected renal cells to quantify the rapid replication of BKV, indicating the progressive nature of PVAN. The authors in [8] extend this work and construct a dynamical model of the BK virus in kidney transplant patients suffering from PVAN. Their results elucidate the relationship between tubular epithelial and urothelial cells with respect to BKV infection. Kepler *et al.* [11] present a model that captures the dynamics of both latent and active HCMV infection in immunosuppressed patients who underwent solid organ transplant (SOT). Banks *et al.* [2] modify and extend this model to include the immune response to the donor kidney, and design an optimal control scheme to find the optimum amount of immunosuppressant for patients who underwent SOT and are susceptible or infected by HCMV [2]. Banks *et al.* [3] modify this model to consider BKV infection and partially validate the resulting model using clinical data. Due to the large number of parameters and limited data, the authors implement an iterative process to identify the most sensitive model parameters to be estimated.

We build on the work of Banks *et al.* [3] and investigate the uncertainty in clinical data from a BKV infected kidney transplant recipient. BKV is the primary etiologic agent in most cases where patients exhibit symptoms of PVAN, although JC virus can also cause PVAN. Reactivation and/or spike in BKV load in renal transplant recipients is often due to the high dosage of the potent immunosuppressants prescribed to decrease the odds of graft rejection. There is currently no approved antiviral drug therapy available to fight BKV; vigilant screening for early detection and monitoring the immunosuppressant dosage are the only prevention methods for symptomatic BKV nephropathy [10].

With this BKV model, we illustrate the use of pseudo-measurement errors and residual plots to detect both mathematical model and statistical model misspecifications. Following [1], we apply second order differencing directly to the data to determine the correct statistical model. After performing an inverse problem with this appropriate statistical model, we demonstrate how residual plots can reveal error in the mathematical model. The remainder of the paper is as follows. Section 2 contains an overview of the BKV model and clinical data. The inverse problem and difference-based methodologies are given in Sections 3 and 4 respectively. Section 5 includes our results using both clinical and simulated data. Lastly, we present our conclusions and plans for future work in Section 6.

2 Mathematical model and clinical data

2.1 Mathematical model synopsis

We consider the following BKV model in [3], which describes the dynamics of the free BK viral load (V), susceptible cells (H_S), BKV infected cells (H_I), BKV-specific CD8+ T cells (E_V), allospecific CD8+ T cells (E_K), and the surrogate for GFR, serum creatinine (C)

$$\dot{H}_S = \lambda_{HS} \left(1 - \frac{H_S}{\kappa_{HS}} \right) H_S - \beta H_S V \quad (1a)$$

$$\dot{H}_I = \beta H_S V - \delta_{HI} H_I - \delta_{EH} E_V H_I \quad (1b)$$

$$\dot{V} = \rho_V \delta_{HI} H_I - \delta_V V - \beta H_S V \quad (1c)$$

$$\dot{E}_V = (1 - \epsilon_I) [\lambda_{EV} + \rho_{EV}(V) E_V] - \delta_{EV} E_V \quad (1d)$$

$$\dot{E}_K = (1 - \epsilon_I) [\lambda_{EK} + \rho_{EK}(H_S) E_K] - \delta_{EK} E_K \quad (1e)$$

$$\dot{C} = \lambda_C - \delta_C(E_K, H_S) C \quad (1f)$$

where

$$\rho_{EV}(V) = \frac{\bar{\rho}_{EV} V}{V + \kappa_V}, \quad (1g)$$

$$\rho_{EK}(H_S) = \frac{\bar{\rho}_{EK} H_S}{H_S + \kappa_{KH}}, \quad (1h)$$

$$\delta_C(E_K, H_S) = \frac{\delta_{C0} \kappa_{EK}}{E_K + \kappa_{EK}} \cdot \frac{H_S}{H_S + \kappa_{CH}}, \quad (1i)$$

and initial conditions,

$$(H_S(0), H_I(0), V(0), E_V(0), E_K(0), C(0)) = (H_{S0}, H_{I0}, V_0, E_{V0}, E_{K0}, C_0). \quad (1j)$$

Susceptible cells proliferate logistically at a maximum rate λ_{HS} with carrying capacity κ_{HS} . Susceptible cells and free virions are both lost due to interacting with each other at rate β , resulting in infected cells. Infected cells lyse at rate δ_{HI} due to the cytopathic effect of the virus and produce ρ_V virions. BKV-specific CD8+ T cells also eliminate infected cells at rate δ_{EH} . The free virus is naturally cleared at rate δ_V . Both the BKV-specific CD8+ T cells and the allospecific CD8+ T cells are inversely related to the immunosuppressant dosage efficiency ϵ_I . The source rates for E_V and E_K are given by λ_{EV} and λ_{EK} respectively. The BKV-specific CD8+ T cells proliferate in the presence of free virions at maximum rate $\bar{\rho}_{EV}$ with half saturation level κ_V . Similarly, the allospecific CD8+ T cells proliferate in the presence of susceptible cells at maximum rate $\bar{\rho}_{EK}$ with half saturation level κ_{KH} . Both E_V and E_K die at constant rates δ_{EV} and δ_{EK} respectively. Creatinine is produced at rate λ_C . The clearance rate for C is dependent upon both E_K and H_S with a maximum clearance rate of δ_{C0} and saturation levels κ_{EK} and κ_{CH} . The efficiency of the immunosuppressant ϵ_I is approximated by the following piecewise constant function

$$\epsilon_I(t) = \begin{cases} \epsilon_1 & t \in [0, 21] \\ \epsilon_2 & t \in (21, 60] \\ \epsilon_3 & t \in (60, 120] \\ \epsilon_4 & t \in (120, 450]. \end{cases} \quad (1k)$$

The state variable descriptions and units can be found in Table 1. The diagrammatic representation of the model (1) is given in Figure 1. Table 2 contains the description of the model parameters and fixed values. See [3] for further details on this model.

Table 1: Description of state variables.

State	Description	Unit
H_S	Concentration of susceptible host cells	cells/mL
H_I	Concentration of infected host cells	cells/mL
V	Concentration of free BKV	copies/mL
E_V	Concentration of BKV-specific CD8+ T cells	cells/mL
E_K	Concentration of allospecific CD8+ T cells that target kidney	cells/mL
C	Concentration of serum creatinine	mg/dL

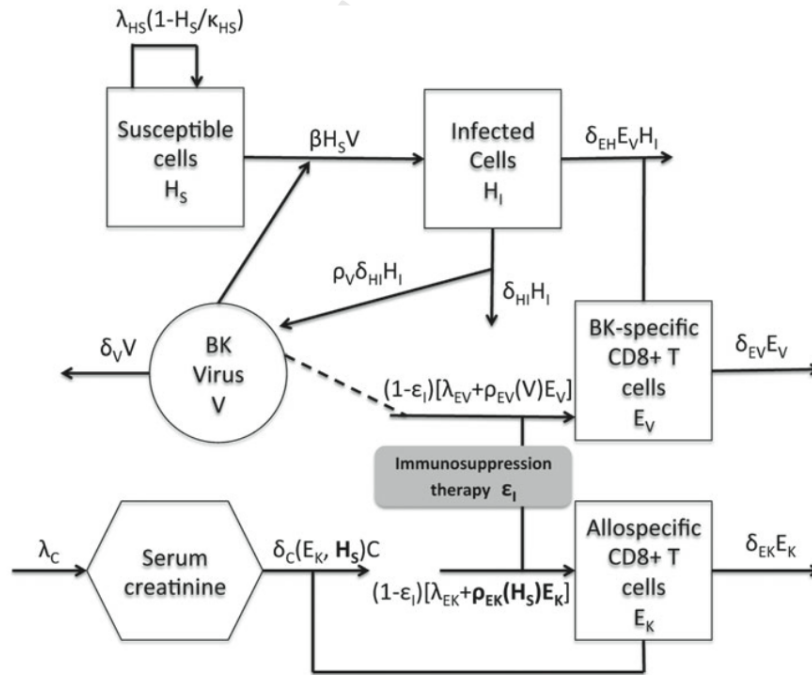


Figure 1: Model diagram of the BKV virus affecting renal cells [3].

Table 2: Model parameter descriptions and fixed values from [3] (est. indicates a free parameter).

Parameter	Description	Unit	Value
λ_{HS}	Proliferation rate for H_S	1/day	0.030
κ_V	Saturation constant	copies/mL	180.676
κ_{HS}	Saturation constant	cells/mL	1025.888
λ_{EK}	Source rate of E_K	cells/(mL·day)	0.002
β	Infection rate of H_S by V	mL/(copies·day)	est.
δ_{EK}	Death rate of E_K	1/day	est.
δ_{HI}	Death rate of H_I by V	1/day	0.085
λ_C	Production rate for C	mg/(dL·day)	0.007
ρ_V	# Virions produced by H_I before death	copies/cells	4292.398
δ_{C0}	Maximum clearance rate for C	1/day	0.014
δ_{EH}	Elimination rate of H_I by E_V	mL/(cells·day)	0.002
κ_{EK}	Saturation constant	cells/mL	0.200
δ_V	Natural clearance rate of V	1/day	0.372
κ_{CH}	Saturation constant	cells/mL	10.000
λ_{EV}	Source rate of E_V	cells/(mL·day)	0.001
$\bar{\rho}_{EK}$	Maximum proliferation rate for E_K	1/day	est.
δ_{EV}	Death rate of E_V	1/day	est.
κ_{KH}	Saturation constant	cells/mL	84.996
$\bar{\rho}_{EV}$	Maximum proliferation rate for E_V	1/day	est.
ϵ_I	Efficacy of immunosuppressive drugs		

2.2 Log-scaled model

Due to a scale difference among model states and model parameters, we use log transformation to resolve any scaling issues during numerical simulations and implementation of the inverse problem (see [3] for details). We can rewrite model (1) as the vector system,

$$\frac{d\mathbf{y}}{dt} = \mathbf{h}(\mathbf{y}, \bar{\mathbf{q}}, \mathbf{y}_0)$$

where

$$\mathbf{y} = [H_S, H_I, V, E_V, E_K, C]^T,$$

$$\bar{\mathbf{q}} = [\lambda_{HS}, \lambda_{EK}, \lambda_{EV}, \lambda_C, \beta, \delta_{EH}, \delta_V, \bar{\rho}_{EV}, \delta_{EV}, \delta_{EK}, \delta_{C0}, \delta_{HI}, \kappa_{CH}, \kappa_{KH}, \kappa_{HS}, \kappa_{EK}, \kappa_V, \rho_V, \bar{\rho}_{EK}, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4]^T,$$

and

$$\mathbf{y}(0) = \mathbf{y}_0.$$

We make the afore-mentioned log transformation by defining variables

$$\begin{aligned} x_i &= \log_{10}(y_i), & i &= 1, 2, 3, 4, 5 \\ x_6 &= y_6, \\ x_{0i} &= \log_{10}(y_{0i}), & i &= 1, 2, 3, 4, 5 \\ x_{06} &= y_{06}, \\ q_j &= \log_{10}(\bar{q}_j), & j &= 1, 2, \dots, 19 \\ q_j &= q_j, & j &= 20, \dots, 23. \end{aligned}$$

Then the log-scaled model becomes

$$\frac{d\mathbf{x}}{dt} = \mathbf{g}(\mathbf{x}, \mathbf{q}, \mathbf{x}_0),$$

where $g_i(\mathbf{x}, \mathbf{q}, \mathbf{x}_0)$ is given by

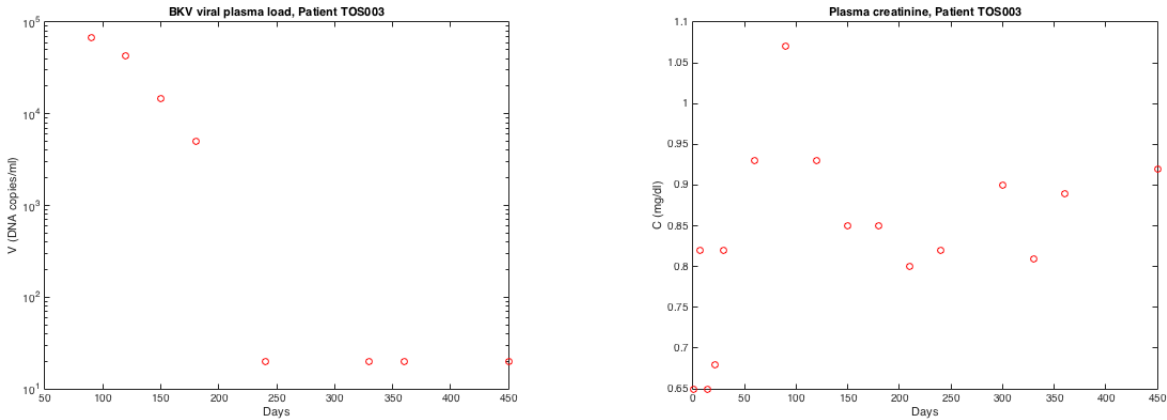
$$\begin{aligned} g_i(\mathbf{x}, \mathbf{q}, \mathbf{x}_0) &= \frac{dx_i}{dt} = \frac{dx_i}{dy_i} \frac{dy_i}{dt} = \frac{1}{y_i \ln(10)} h_i(\mathbf{y}, \bar{\mathbf{q}}, \mathbf{y}_0), \quad i = 1, 2, 3, 4, 5 \\ &= \frac{1}{10^{x_i} \ln(10)} h_i(10^{(x_1, x_2, \dots, x_5)}, x_6, 10^{(q_1, q_2, \dots, q_{19})}, q_{20}, q_{21}, \dots, q_{23}, 10^{(x_{01}, x_{02}, \dots, x_{05})}, x_{06}), \end{aligned}$$

and

$$\begin{aligned} g_6(\mathbf{x}, \mathbf{q}, \mathbf{x}_0) &= h_6(\mathbf{y}, \bar{\mathbf{q}}, \mathbf{y}_0) \\ &= h_6(10^{(x_1, x_2, \dots, x_5)}, x_6, 10^{(q_1, q_2, \dots, q_{19})}, q_{20}, q_{21}, \dots, q_{23}, 10^{(x_{01}, x_{02}, \dots, x_{05})}, x_{06}). \end{aligned}$$

2.3 Clinical data

We investigate uncertainty in the clinical data [3]. This data set consists of eight BK viral plasma load (DNA copies/mL) measurements and sixteen plasma creatinine level (mg/dL) measurements for patient TOS003 from Massachusetts General Hospital. The patient was diagnosed with BKV infection in the first 3 months of transplantation. With every visit, dosage and combination of immunosuppressants were updated. Figure 2 contains the plots of the data.



(a) BK viral load data

(b) Creatinine data

Figure 2: Patient TOS003 BKV viral plasma loads and plasma creatinine levels.

3 Inverse problem method and statistical model selection

We follow standard inverse problem procedures to estimate parameters in our mathematical model [4, 5, 7, 14]. Consider a general N -dimensional dynamical system with parameter vector \mathbf{q} ,

$$\begin{aligned} \frac{d\mathbf{x}}{dt}(t) &= \mathbf{g}(t, \mathbf{x}(t); \mathbf{q}), \\ \mathbf{x}(t_0) &= \mathbf{x}_0, \end{aligned}$$

with an m dimensional observation process

$$f(t; \boldsymbol{\theta}) = \mathcal{C}\mathbf{x}(t; \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\mathbf{q}^\top, \mathbf{x}_0^\top)^\top$ is the vector of parameters along with the initial conditions to be estimated and \mathcal{C} is the $m \times N$ observation matrix.

Our data set consists of observed values for the plasma viral load and creatinine levels. Thus our observation matrix is the following

$$\mathcal{C} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

as $\mathbf{x} = [\log_{10} H_S, \log_{10} H_I, \log_{10} V, \log_{10} E_V, \log_{10} E_K, C]^\top$ and $\mathbf{f} = [\log_{10} V, C]^\top$.

Let y_i^1 represent the free BK viral load measurements and y_j^2 represent the plasma creatinine load measurements at time points $t_i^1, i = 1, 2, \dots, n_1$ and $t_j^2, j = 1, 2, \dots, n_2$ respectively. Here $n_1 = 8$ and $n_2 = 16$. We note that there is some discrepancy between the actual phenomenon, which is represented through the data, and the above observation process. We account for this uncertainty with the following statistical model,

$$\begin{aligned} Y_i^1 &= f_1(t_i^1; \boldsymbol{\theta}_0) + f_1(t_i^1; \boldsymbol{\theta}_0)^{\gamma_1} \mathcal{E}_i^1, & i = 1, 2, \dots, n_1, \\ Y_j^2 &= f_2(t_j^2; \boldsymbol{\theta}_0) + f_2(t_j^2; \boldsymbol{\theta}_0)^{\gamma_2} \mathcal{E}_j^2, & j = 1, 2, \dots, n_2, \end{aligned}$$

where $\boldsymbol{\gamma} \geq \mathbf{0}$ and the $p \times 1$ vector $\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}$ is the ‘‘true’’ or nominal parameter set. Here $f_1(t_i^1; \boldsymbol{\theta}_0) = x_3(t_i^1; \boldsymbol{\theta}_0)$ and $f_2(t_j^2; \boldsymbol{\theta}_0) = x_6(t_j^2; \boldsymbol{\theta}_0)$. The $n_1 \times 1$ and $n_2 \times 1$ random error vectors \mathcal{E}_i^1 and \mathcal{E}_j^2 respectively are assumed to be independent and identically distributed (*i.i.d*) with mean zero and $\text{Var}(\mathcal{E}_i^1) = \sigma_{01}^2$ and $\text{Var}(\mathcal{E}_j^2) = \sigma_{02}^2$. The corresponding realizations are,

$$\begin{aligned} \mathbf{y}_i^1 &= f_1(t_i^1; \boldsymbol{\theta}_0) + f_1(t_i^1; \boldsymbol{\theta}_0)^{\gamma_1} \epsilon_i^1, & i = 1, 2, \dots, n_1, \\ \mathbf{y}_j^2 &= f_2(t_j^2; \boldsymbol{\theta}_0) + f_2(t_j^2; \boldsymbol{\theta}_0)^{\gamma_2} \epsilon_j^2, & j = 1, 2, \dots, n_2. \end{aligned}$$

For $\boldsymbol{\gamma} \geq \mathbf{0}$, a generalized least squares method is appropriate to perform the inverse problem. In order to estimate $\boldsymbol{\theta}_0$, we want to minimize the distance between the collected data and mathematical model, where the observables are weighted according to their variability and, for each observable, the observations over time are weighted unequally.

We first approximate σ_{01}^2 and σ_{02}^2 by the following

$$\begin{aligned} \hat{\sigma}_{01}^2 &= \frac{1}{n_1 - p} \sum_{i=1}^{n_1} \left(\frac{y_i^1 - f_1(t_i^1; \hat{\boldsymbol{\theta}}_{GLS})}{f_1(t_i^1; \hat{\boldsymbol{\theta}}_{GLS})^{\gamma_1}} \right)^2 \\ \hat{\sigma}_{02}^2 &= \frac{1}{n_2 - p} \sum_{j=1}^{n_2} \left(\frac{y_j^2 - f_2(t_j^2; \hat{\boldsymbol{\theta}}_{GLS})}{f_2(t_j^2; \hat{\boldsymbol{\theta}}_{GLS})^{\gamma_2}} \right)^2. \end{aligned}$$

Then, we solve iteratively the following system of equations to numerically determine $\hat{\boldsymbol{\theta}}_{GLS}$:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{GLS} &= \underset{\boldsymbol{\theta} \in \boldsymbol{\Omega}}{\text{argmin}} \left(\sum_{i=1}^{n_1} [y_i^1 - f_1(t_i^1; \boldsymbol{\theta})]^\top \hat{V}_1^{-1}(t_i^1) [y_i^1 - f_1(t_i^1; \boldsymbol{\theta})] \right. \\ &\quad \left. + \sum_{j=1}^{n_2} [y_j^2 - f_2(t_j^2; \boldsymbol{\theta})]^\top \hat{V}_2^{-1}(t_j^2) [y_j^2 - f_2(t_j^2; \boldsymbol{\theta})] \right) \end{aligned} \quad (2)$$

$$\hat{V}_1(t_i^1) = \frac{f_1(t_i^1; \hat{\boldsymbol{\theta}}_{GLS})^{2\gamma_1}}{n_1 - p} \sum_{i=1}^{n_1} \left(\frac{y_i^1 - f_1(t_i^1; \hat{\boldsymbol{\theta}}_{GLS})}{f_1(t_i^1; \hat{\boldsymbol{\theta}}_{GLS})^{\gamma_1}} \right)^2 \quad (3)$$

$$\hat{V}_2(t_j^2) = \frac{f_2(t_j^2; \hat{\boldsymbol{\theta}}_{GLS})^{2\gamma_2}}{n_2 - p} \sum_{j=1}^{n_2} \left(\frac{y_j^2 - f_2(t_j^2; \hat{\boldsymbol{\theta}}_{GLS})}{f_2(t_j^2; \hat{\boldsymbol{\theta}}_{GLS})^{\gamma_2}} \right)^2. \quad (4)$$

We use the following iterative procedure [4, 5, 7, 14] :

1. Estimate $\hat{\boldsymbol{\theta}}_{GLS}^{(0)}$ using (2) with $\hat{V}_1(t_i^1) = 1$ and $\hat{V}_2(t_j^2) = 1$. Set $k = 0$.
2. Compute weights $\hat{\omega}_i^1 = f_1(t_i^1, \hat{\boldsymbol{\theta}}_{GLS}^{(k)})^{2\gamma_1}$ and $\hat{\omega}_j^2 = f_2(t_j^2, \hat{\boldsymbol{\theta}}_{GLS}^{(k)})^{2\gamma_2}$.
3. Solve for $\hat{V}_1(t_i^1)^{(k)}$ and $\hat{V}_2(t_j^2)^{(k)}$ using $\hat{\boldsymbol{\theta}}_{GLS}^{(k)}$, $\hat{\omega}_i^1$, and $\hat{\omega}_j^2$ in equations (3) and (4) respectively.
4. Estimate $\hat{\boldsymbol{\theta}}_{GLS}^{(k+1)}$ using $\hat{V}_1(t_i^1)^{(k)}$ and $\hat{V}_2(t_j^2)^{(k)}$ in equation (2).
5. Set $k := k + 1$ and return to step 2. Terminate when two successive estimates for $\hat{\boldsymbol{\theta}}_{GLS}$ are sufficiently close.

Note that this is not the same as taking the derivative of the right hand side of (2) and setting it equal to zero. For more details see page 33 of [4] and page 89 of [14].

If we assume $\boldsymbol{\gamma} = [0, 0]$, then our statistical model is called an absolute error model and an ordinary least squares method is appropriate for parameter estimation. Banks *et al.* [3] consider an absolute error model and additionally assume that the variances for each observable are equal (i.e., $\sigma_{01}^2 = \sigma_{02}^2$). While the statistical model choice in [3] yields good results, we believe it is more biologically realistic to assume the variance in observation errors are not equal and the size of the observation error is proportional to the size of the observed quantity.

4 Difference-based methods and residuals

We use a second order difference-based method to determine the correct statistical model ($\boldsymbol{\gamma}$ value) [1]. Another method often implemented consists of performing an inverse problem with some $\boldsymbol{\gamma}$ value and computing the residuals

$$r_{kl} = \frac{y_l^k - f_k(t_l^k; \hat{\boldsymbol{\theta}})}{f_k(t_l^k; \hat{\boldsymbol{\theta}})^{\gamma_k}}, \quad (5)$$

for each observable k at time $t_l, l = 1, \dots, n_k$. The plots of r_{kl} vs. t_l should be randomly scattered around the x -axis. If an undesired megaphone shape is present, then a different $\boldsymbol{\gamma}$ value is chosen and the process is repeated until a $\boldsymbol{\gamma}$ value produces the desired scatter plot. However, this method does not consider both the mathematical model and statistical model misspecifications; it determines the correct statistical model under the tacit assumption that one has a correct mathematical model. It is also time consuming as it might take several attempts of performing an inverse problem and plotting the residuals until a good statistical model is chosen.

We follow [1] and first apply the second order difference-based method directly to the data to determine the correct $\boldsymbol{\gamma}$ value, which is both computationally economical as well as time efficient

and independent of any assumed correct mathematical model. We first calculate the following pseudo measurement errors for observable k at time $t_l, l = 1, \dots, n_k$

$$\hat{\epsilon}_l^k = \begin{cases} \frac{1}{\sqrt{2}}(y_{l+1}^k - y_l^k) & \text{for } l = 1 \\ \frac{1}{\sqrt{6}}(y_{l-1}^k - 2y_l^k + y_{l+1}^k) & \text{for } l = 2, \dots, n_k-1 \\ \frac{1}{\sqrt{2}}(y_l^k - y_{l-1}^k) & \text{for } l = n_k. \end{cases}$$

Next we calculate the modified residuals

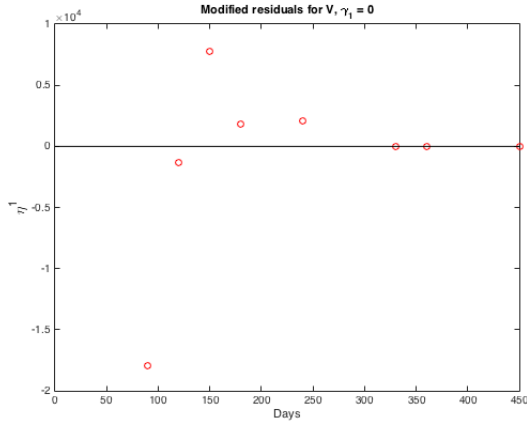
$$\eta_l^k = \frac{\hat{\epsilon}_l^k}{|y_l^k - \hat{\epsilon}_l^k|^{\gamma_k}}$$

for observable k at time $t_l, l = 1, \dots, n_k$ for different values of γ_k . We plot these modified residuals vs. time for different γ_k values to find the γ_k value that produces a random scatter plot. Once the correct observational error is accounted for, we perform the inverse problem with this statistical model and compute the residuals in (5). If the residual plots are not randomly distributed around the x -axis, then the error must be due to mathematical model misspecification, implying another iteration of the modeling process is needed.

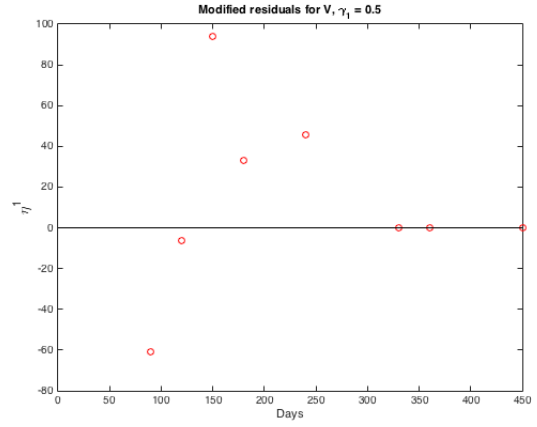
5 Results

5.1 Clinical Data

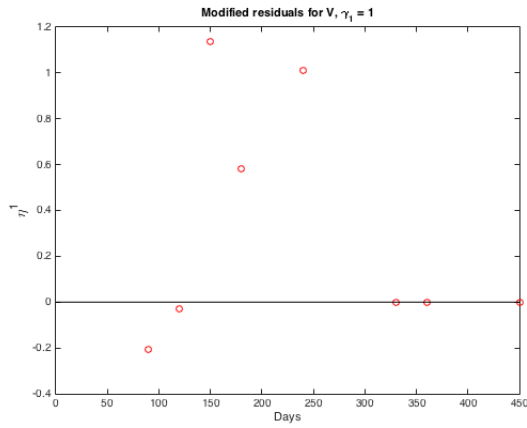
Using second order differencing, we plot the modified residuals for both viral load and creatinine versus time and visually assess the plots to choose an appropriate γ value. Figure 3 contains the graphs of the viral load modified residuals vs. time for various γ_1 values. As can be seen due to the limited amount of data, it is difficult to determine the correct γ_1 value through visual assessment. The value $\gamma_1 = 0.5$ provides an approximately symmetric distribution around the x axis with relatively small residual values. The creatinine modified residuals for different γ_2 values are given in Figure 4. The modified residuals with $\gamma_2 = 0$ appear to be randomly distributed whereas the modified residuals with $\gamma_2 = 1$ reveal a slight non-random (megaphone) shape. Even though we visually assess the plots to pick a suitable γ value to the best of our ability, the sparseness of the data set makes it difficult to make a stronger case for a particular statistical model.



(a) $\gamma_1 = 0$

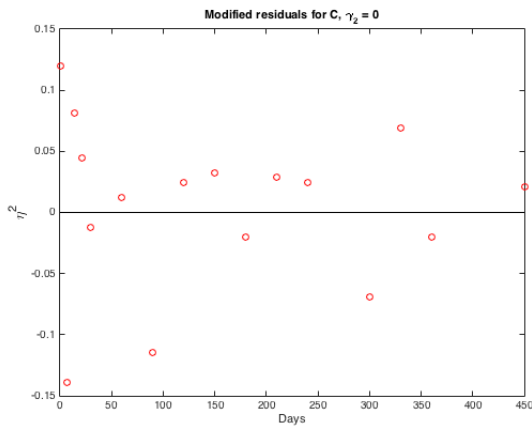


(b) $\gamma_1 = 0.5$

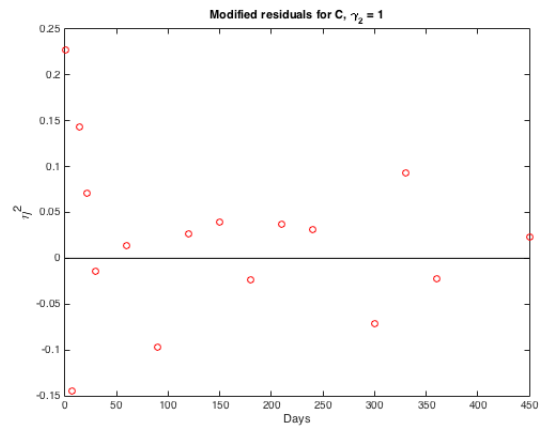


(c) $\gamma_1 = 1$

Figure 3: Viral load modified residuals vs. time for various γ_1 values.



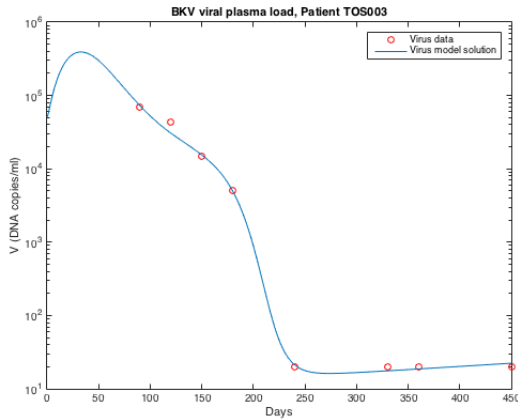
(a) $\gamma_2 = 0$



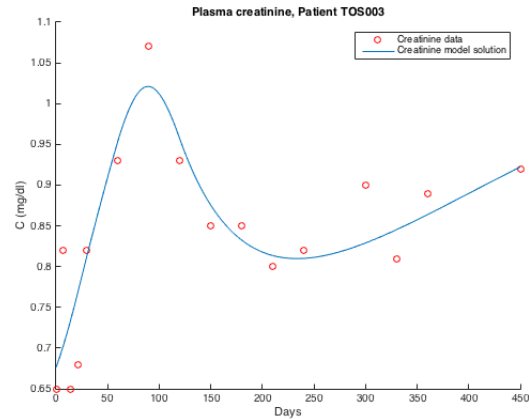
(b) $\gamma_2 = 1$

Figure 4: Creatinine modified residuals vs. time for various γ_2 values.

With our best guess of the correct statistical model, $\gamma = [0.5, 0]$, we next perform an inverse problem for the 5 most sensitive parameters [3] and obtain the residuals in order to detect the presence of mathematical model error. We perform the inverse problem using the inbuilt MATLAB function `fmincon`. The initial guesses for the parameters are those used in [3] and lower and upper bounds are set for each of the 5 parameters for computational efficiency. We can see from Figure 5 that the model solution fits the data well and the corresponding residuals in Figure 6 appear to form a random band around the x -axis. This suggests that the mathematical model accurately describes the biological process, although again it is difficult to conclude this with conviction due to the limited amount of data.

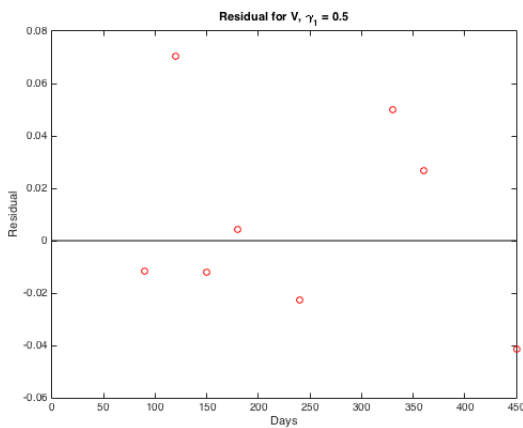


(a) BK virus model solution and data

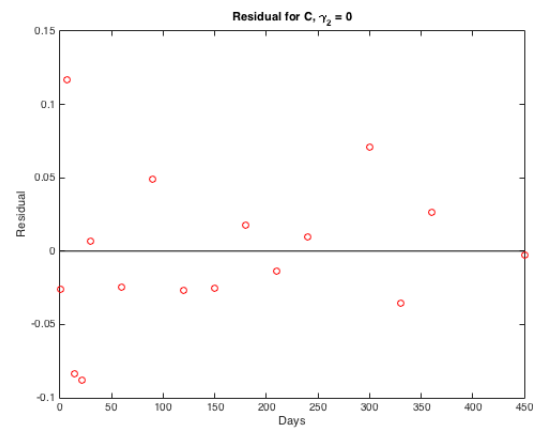


(b) Creatinine model solution and data

Figure 5: Model (1) solution and clinical data with $\gamma = [0.5, 0]$ and $[\log_{10} \beta, \log_{10} \bar{\rho}_{EV}, \log_{10} \delta_{EV}, \log_{10} \delta_{EK}, \log_{10} \bar{\rho}_{EK}] = [-7.061, -0.632, -1.007, -0.92, -0.704]$.



(a) Residuals for BK virus



(b) Residuals for creatinine

Figure 6: Residuals for V and C with $\gamma = [0.5, 0]$.

5.2 Simulated data

When using the second order difference-based method with sparse clinical data, it is not very easy to pick a statistical model or make a strong case for the presence/absence of mathematical model misspecification. To illustrate the need for a denser data set, we repeat the above process with the following simulated data created by adding noise to the “true” model solution

$$V_i = f_1(t_i^1; \boldsymbol{\theta}_0) + f_1(t_i^1; \boldsymbol{\theta}_0)^{\gamma_1} \epsilon_i^1, \quad (6a)$$

$$C_j = f_2(t_j^2; \boldsymbol{\theta}_0) + f_2(t_j^2; \boldsymbol{\theta}_0)^{\gamma_2} \epsilon_j^2, \quad (6b)$$

where $\boldsymbol{\epsilon}^1 \sim N(0, 0.3)$, $\boldsymbol{\epsilon}^2 \sim N(0, 0.03)$, $\boldsymbol{\gamma} = [0.5, 0]$, and the estimated parameters $[\log_{10} \beta, \log_{10} \bar{\rho}_{EV}, \log_{10} \delta_{EV}, \log_{10} \delta_{EK}, \log_{10} \bar{\rho}_{EK}] = [-7.067, -0.601, -0.964, -0.995, -0.785]$. We assume that “data” is collected every week for $\mathbf{t}^1 = \mathbf{t}^2 = [0, 7, 14, \dots, 448]$. As expected, the second order differencing method produces the desired scatter plot for $\boldsymbol{\gamma} = [0.5, 0]$ and undesired megaphone shapes for other $\boldsymbol{\gamma}$ values. The residual plots also exhibit no mathematical model misspecification, which is expected since the data was created using the mathematical model (see Appendix A).

We now demonstrate how the residual plots can detect mathematical model error or misspecification by performing an inverse problem with a simpler version of model (1), given in (7). While the original model (1) assumes the susceptible population grows logistically, the simpler model (7) assumes a growth rate of $\lambda_{HS} - \delta_{HS}H_S$, where H_S cells are produced at a constant rate λ_{HS} and die at a rate δ_{HS} . The simpler model is given by the following

$$\dot{H}_S = \lambda_{HS} - \delta_{HS}H_S - \beta H_S V \quad (7a)$$

$$\dot{H}_I = \beta H_S V - \delta_{HI}H_I - \delta_{EH}E_V H_I \quad (7b)$$

$$\dot{V} = \rho_V \delta_{HI}H_I - \delta_V V - \beta H_S V \quad (7c)$$

$$\dot{E}_V = (1 - \epsilon_I)[\lambda_{EV} + \rho_{EV}(V)E_V] - \delta_{EV}E_V \quad (7d)$$

$$\dot{E}_K = (1 - \epsilon_I)[\lambda_{EK} + \rho_{EK}(H_S)E_K] - \delta_{EK}E_K \quad (7e)$$

$$\dot{C} = \lambda_C - \delta_C(E_K, H_S)C \quad (7f)$$

where

$$\rho_{EV}(V) = \frac{\bar{\rho}_{EV}V}{V + \kappa_V}, \quad (7g)$$

$$\rho_{EK}(H_S) = \frac{\bar{\rho}_{EK}H_S}{H_S + \kappa_{KH}}, \quad (7h)$$

$$\delta_C(E_K, H_S) = \frac{\delta_{C0}\kappa_{EK}}{E_K + \kappa_{EK}} \cdot \frac{H_S}{H_S + \kappa_{CH}}, \quad (7i)$$

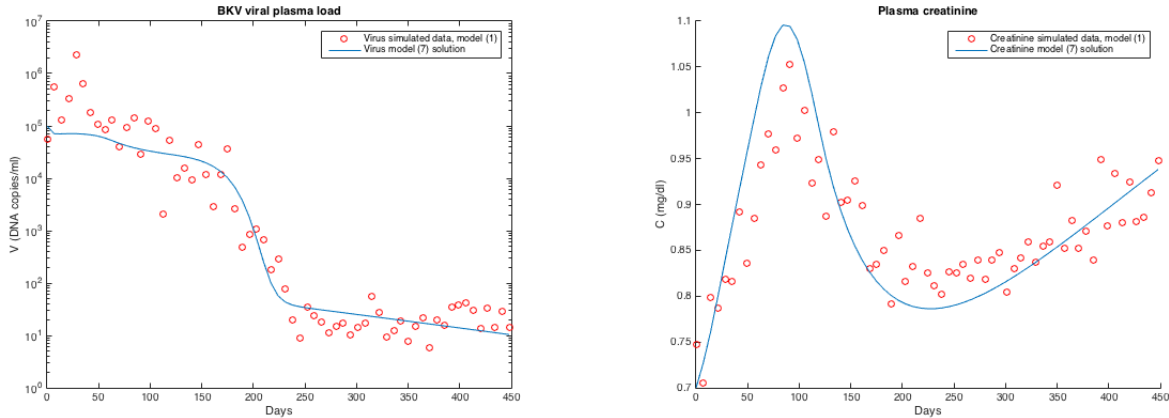
and initial conditions,

$$(H_S(0), H_I(0), V(0), E_V(0), E_K(0), C(0)) = (H_{S0}, H_{I0}, V_0, E_{V0}, E_{K0}, C_0). \quad (7j)$$

The immunosuppressant efficiency is defined by the piecewise constant function (1k).

We perform the inverse problem using $\boldsymbol{\gamma} = [0.5, 0]$ to estimate the 6 parameters β , $\bar{\rho}_{EV}$, δ_{EV} , δ_{EK} , δ_{HS} , $\bar{\rho}_{EK}$ and obtain the solutions in Figure 7. Even though the simpler model produces a

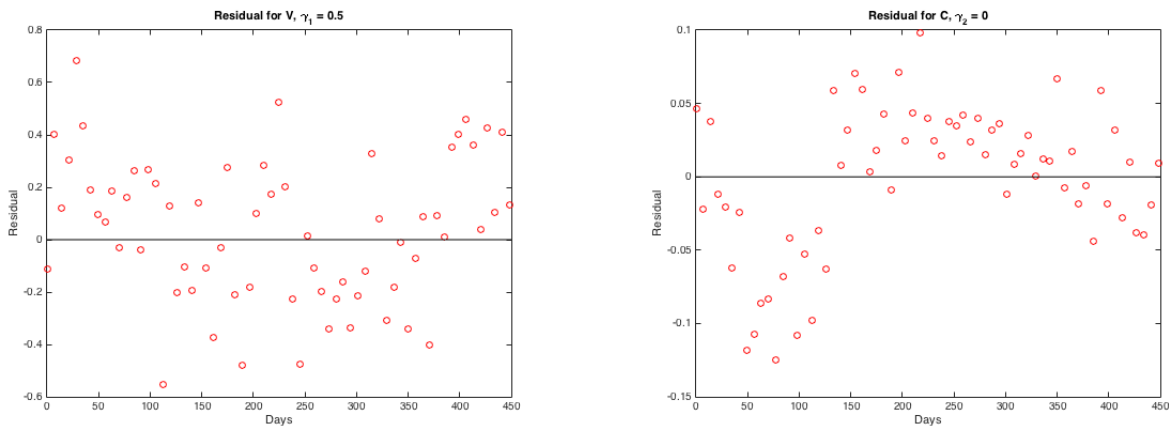
reasonable fit to the data, the residuals produce a strong non-random pattern (Figure 8). Since we already eliminated statistical error model discrepancy (through the difference-based method), these non-random residuals indicate a mathematical model misspecification. That is, the simpler model (7) is unable to accurately capture the dynamics represented in the data.



(a) BK virus model solution and simulated data

(b) Creatinine model solution and simulated data

Figure 7: Model (7) solution and simulated data created from model (1) with $\gamma = [0.5, 0]$ and $[\log_{10} \beta, \log_{10} \bar{\rho}_{EV}, \log_{10} \delta_{EV}, \log_{10} \delta_{EK}, \log_{10} \delta_{HS}, \log_{10} \bar{\rho}_{EK}] = [-7.747, -0.398, -0.710, -0.690, -4.600, -0.492]$.



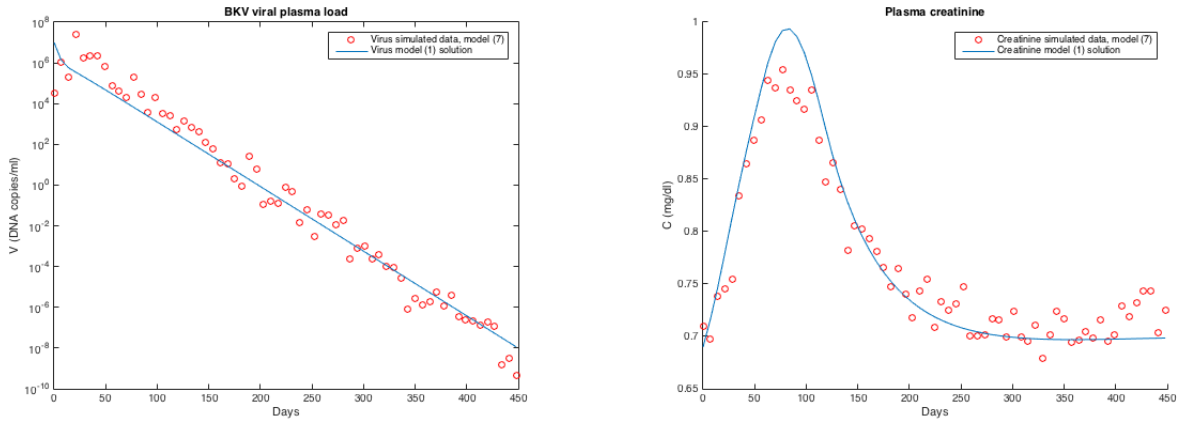
(a) Residuals for BK virus

(b) Residuals for creatinine

Figure 8: Residuals for V and C with $\gamma = [0.5, 0]$.

We next investigate mathematical model misspecification when a more complex model than warranted is assumed. To do so, we create a new simulated data set for $\mathbf{t}^1 = \mathbf{t}^2 = [0, 7, 14, \dots, 448]$ using (6) where \mathbf{f}_1 and \mathbf{f}_2 now represent $\log_{10} V$ and C in model (7), $\boldsymbol{\gamma} = [0, 0.7]$, $\boldsymbol{\mathcal{E}}^1 \sim N(0, 0.5)$, and $\boldsymbol{\mathcal{E}}^2 \sim N(0, 0.02)$. Parameter values from Table 2 were used to create the new simulated data set with free parameters $[\log_{10} \beta, \log_{10} \bar{\rho}_{EV}, \log_{10} \delta_{EV}, \log_{10} \delta_{EK}, \log_{10} \bar{\rho}_{EK}] = [-7.067, -0.601, -0.964, -0.995, -0.785]$ and additional parameter $\delta_{HS} = 0.003$ copies/day [8].

As expected, the difference-based method with $\boldsymbol{\gamma} = [0, 0.7]$ produces random scatter plots (see Appendix B). We perform the inverse problem with this data set and the original model (1). The model (1) solutions and corresponding residuals are plotted in Figure 9 and Figure 10. Even though the fit between the model and data looks acceptable, the residuals display a strong pattern, indicating incorrect mathematical model assumptions.

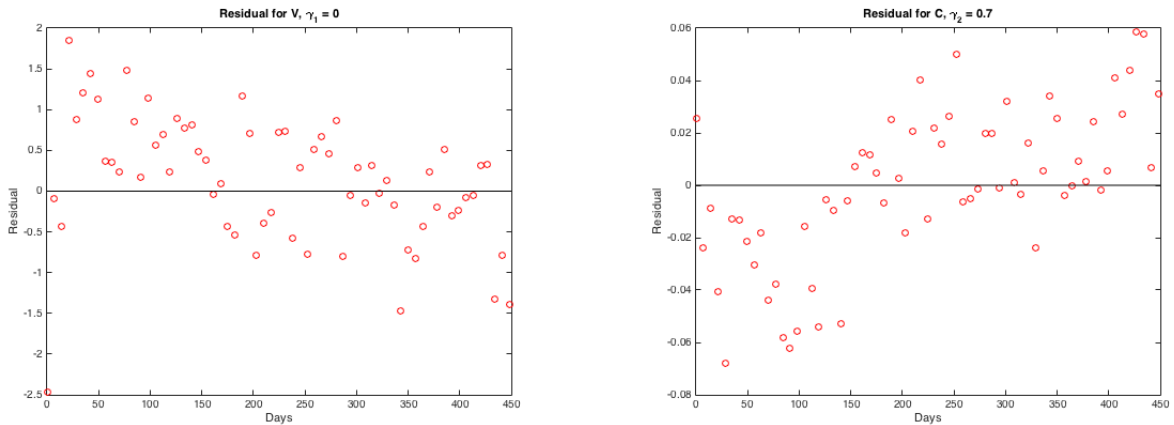


(a) BK virus model solution and simulated data

(b) Creatinine model solution and simulated data

Figure 9: Inverse problem model solution with (1) and simulated data with (7) with $\boldsymbol{\gamma} = [0, 0.7]$ and

$$[\log_{10} \beta, \log_{10} \bar{\rho}_{EV}, \log_{10} \delta_{EV}, \log_{10} \delta_{EK}, \log_{10} \bar{\rho}_{EK}] = [-8.020, -0.744, -0.223, -0.863, -0.675].$$



(a) Residuals for BK virus

(b) Residuals for creatinine

Figure 10: Residuals for V and C with $\boldsymbol{\gamma} = [0, 0.7]$.

6 Conclusion

We investigated mathematical model and statistical model misspecifications in the context of least squares methodology using a BKV model and both clinical and simulated data. Banks *et al.* [3] use ordinary least squares techniques to perform an inverse problem with clinical data. We build on this work and assume what we believe is a more biologically realistic statistical error model; we consider different variances for different observables and allow the error to depend on the size of the observables (measurements). We follow [1] and demonstrate how difference-based methods can be applied directly to data to determine the correct statistical model and further, we illustrate the use of residuals to detect mathematical model discrepancy. The presence of mathematical model error suggests possibly another iteration of modeling might be needed. However, due to the limited amount of clinical data, no strong conclusion can be reached.

We thereby demonstrate these methods using dense simulated data. We first create data using the BKV model (1) with an associated statistical model. The difference-based method correctly identifies the assumed γ value. Using this statistical model, we perform an inverse problem using a simpler model (7) where the nonlinearity is removed from the susceptible cell population growth dynamics. While the model (7) solution fits the data reasonably well, the residual plots depict a strong pattern, identifying the mathematical model discrepancy. We then repeat the process by creating data using the simpler model (7) and perform an inverse problem with the original model (1). That is, we now assume a more complex dynamical system in comparison to the biological process represented by the data. Again, the residuals indicate error in the mathematical model. Therefore this method can reveal mathematical model misspecification when either simpler or more complex models are assumed as compared to the data dynamics.

Previously, residual plots were solely used to determine the correct statistical model by iteratively performing multiple inverse problems until the correct statistical model was chosen [4]. Using both the difference-based method as well as residual plots is computationally more efficient; the difference-based method can be applied directly to the data and thus multiple inverse problems need not be performed. However, and even more notably, the previous method (using only residual plots) determines the statistical model under the uninformed assumption of a correct mathematical model; the use of both the difference-based method and residuals accounts for both types of error in the inverse problem without prior model assumptions.

Future work includes development of feedback control methodology to develop an adaptive immunosuppressant treatment schedule to balance under- and over-suppression of the immune system for individual renal (and possible other) transplant recipients. However, our results here demonstrate that more data is needed in order to verify that the correct mathematical model is assumed for the control problem.

Acknowledgements

This research was supported in part by the National Institute on Alcohol Abuse and Alcoholism under grant number 1R01AA022714-01A1, and in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-15-1-0298.

References

- [1] H.T. Banks, J. Catenacci, and S. Hu, Use of difference-based methods to explore statistical and mathematical model discrepancy in inverse problems, *Journal of Inverse and Ill-posed Problems*, **24** (2016), 413–433.
- [2] H.T. Banks, S. Hu, T. Jang, and H.D. Kwon, Modeling and optimal control of immune response of renal transplant recipients, *Journal of Biological Dynamics*, **6** (2012), 539–567.
- [3] H.T. Banks, S. Hu, K. Link, E.S. Rosenberg, S. Mitsuma, and L. Rosario, Modeling immune response to BK virus infection and donor kidney in renal transplant recipients, *Inverse Problems in Science and Engineering*, **24** (2016), 127–152.
- [4] H.T. Banks, S. Hu, and W.C. Thompson, *Modeling and Inverse Problems in the Presence of Uncertainty*, Taylor/Francis-Chapman/Hall-CRC Press, Boca Raton, FL, 2014.
- [5] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, Taylor/Francis-Chapman/Hall-CRC Press, Boca Raton, FL, 2009.
- [6] D.L. Bohl and D.C. Brennan, BK virus nephropathy and kidney transplantation, *Clinical Journal of the American Society of Nephrology*, **2** (2007), S36–S46.
- [7] M. Davidian and D.M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London, 2000.
- [8] G.A. Funk, R. Gosert, P. Comoli, F. Ginevri, and H.H. Hirsch, Polyomavirus BK replication dynamics in vivo and in silico to predict cytopathology and viral clearance in kidney transplants, *Am. J. Transplant*, **8** (2008), 2368–2377.
- [9] G.A. Funk, J. Steiger, H.H. Hirsch, Rapid dynamics of polyomavirus type BK in renal transplant recipients, *J. Infect. Dis.*, **190** (2006), 80–87.
- [10] H.H. Hirsch, C.B. Drachenberg, J. Steiger, and E. Ramos, Polyomavirus-associated nephropathy in renal transplantation: critical issues of screening and management, *Advances in Experimental Medicine and Biology*, **577** (2006), 160–173.
- [11] G.M. Kepler, H.T. Banks, M. Davidian, and E.S. Rosenberg, A model for HCMV infection in immunosuppressed patients, *Mathematical and Computational Modelling*, **49** (2009), 1653–1663.
- [12] National Kidney Foundation, <https://www.kidney.org>.
- [13] OPTN/SRTR 2012 Annual Data Report: Kidney.
- [14] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, J. Wiley & Sons, Hoboken, NJ, 2003.
- [15] J. Sellarés, D.G. de Freitas, M. Mengel, J. Reeve, G. Einecke, B. Sis, L.G. Hidalgo, K. Famulski, A. Matas, and P.F. Halloran, Understanding the causes of kidney transplant failure: the dominant role of antibody-mediated rejection and nonadherence, *American Journal of Transplantation*, **12** (2012), 388–399.
- [16] 2015 USRDS Annual Data Report Volume 2: ESRD in the United States.

A Simulated data from the original model (1)

We apply the difference-based method to the simulated data set (6) created using the original model (1) to determine the correct γ value. The modified residuals for the viral load with various γ_1 values are given in Figure 11. As expected, $\gamma_1 = 0.5$ produces the desired scatter plot whereas $\gamma_1 = 0$ and $\gamma_1 = 1$ produce undesired megaphone shapes. Figure 12a contains the desired modified residuals for creatinine levels with $\gamma_2 = 0$. While Figure 12b with $\gamma_2 = 1.5$ is similar, it is not quite as symmetric and has larger modified residual values.

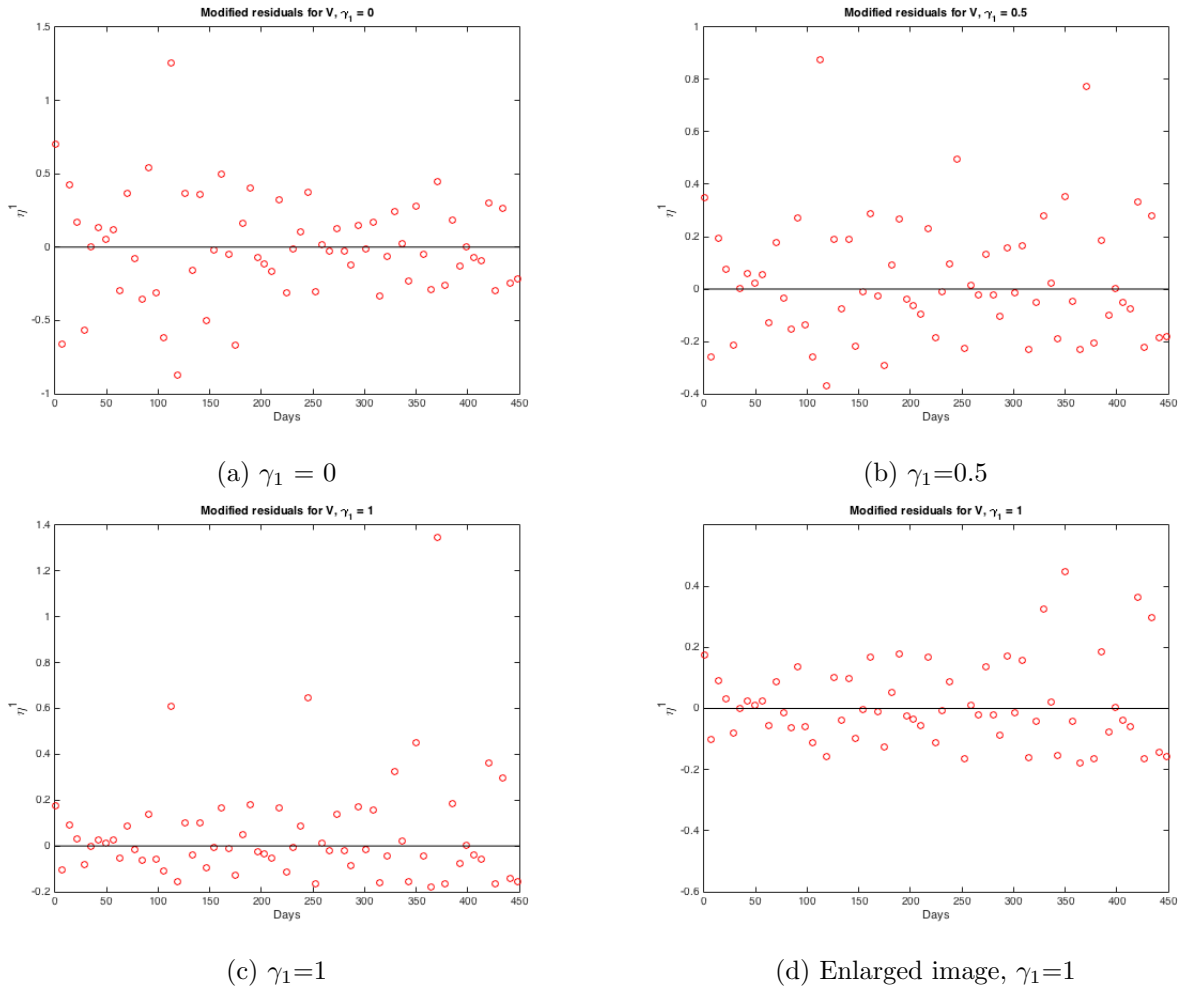
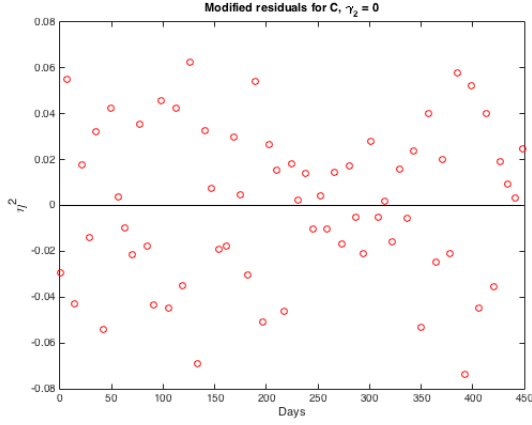
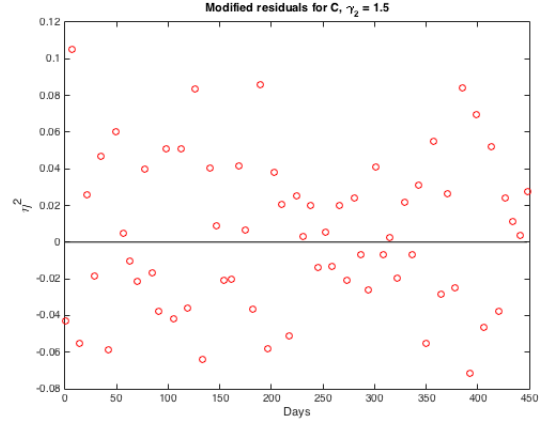


Figure 11: Simulated viral load modified residuals vs. time for various γ_1 values.



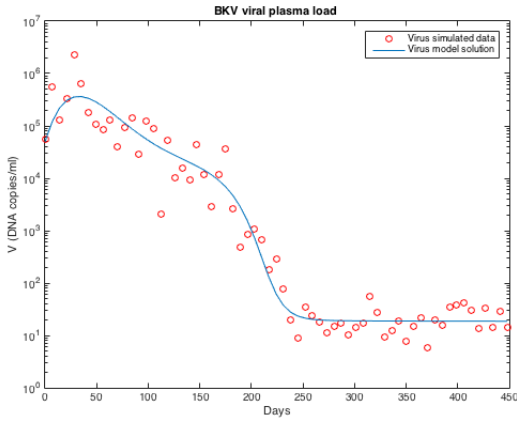
(a) $\gamma_2 = 0$



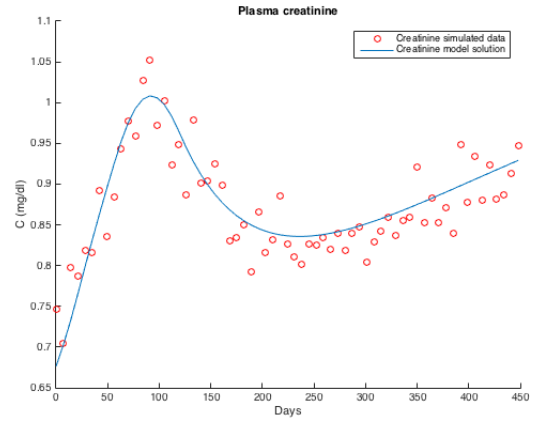
(b) $\gamma_2 = 1.5$

Figure 12: Simulated creatinine modified residuals vs. time for various γ_2 values.

We then solve the inverse problem using the original model (1) with $\boldsymbol{\gamma} = [0.5, 0]$ and plot the residuals to verify there is no mathematical model misspecification. The model solutions and corresponding residuals are plotted in Figure 13 and Figure 14. As expected, the model solutions fit the data well and the corresponding residuals appear to form a uniform band around the x -axis.

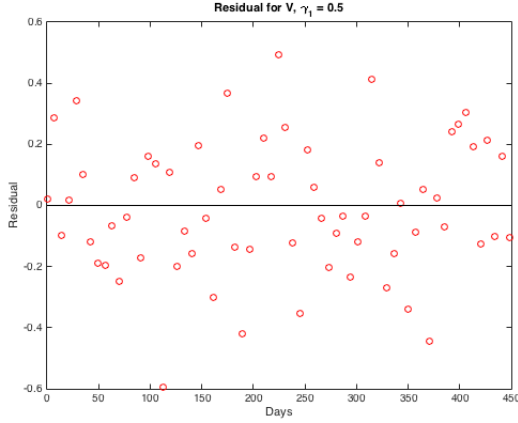


(a) V model solution and simulated data

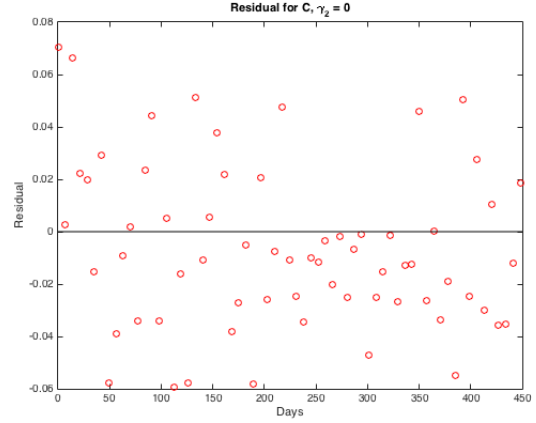


(b) C model solution and simulated data

Figure 13: Inverse problem model (1) solution and simulated data with $\boldsymbol{\gamma} = [0.5, 0]$ and $[\log_{10} \beta, \log_{10} \bar{\rho}_{EV}, \log_{10} \delta_{EV}, \log_{10} \delta_{EK}, \log_{10} \bar{\rho}_{EK}] = [7.0735, -0.6008, -0.9628, -0.9948, -0.7836]$.



(a) Residuals for V

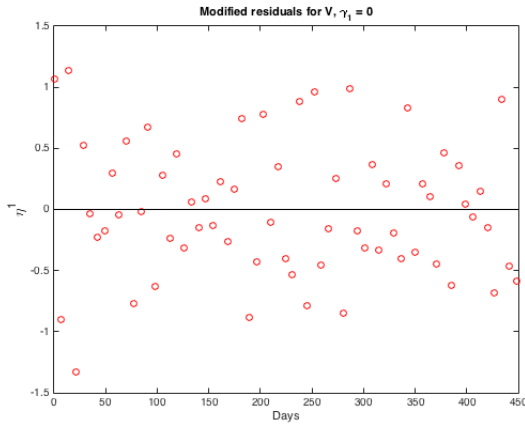


(b) Residuals for C

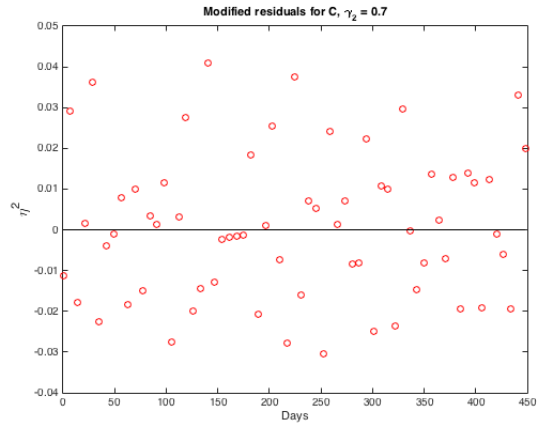
Figure 14: Residuals for V and C with $\gamma = [0.5, 0]$.

B Simulated data from the simpler model (7)

We apply the difference-based method to the simulated data set (6) generated using the simpler model (7) to determine the correct γ value. The randomness in the modified residuals for both the viral load with $\gamma_1 = 0$ (Figure 15a) and the creatinine levels with $\gamma_2 = 0.7$ (Figure 15b) reiterate that the difference-based method works as expected.



(a) Viral load, $\gamma_1 = 0$



(b) Creatinine, $\gamma_2 = 0.7$

Figure 15: Modified residuals for the viral load and creatinine $\gamma = [0, 0.7]$.