Transnational University Limburg is a unique collaboration of two universities in two countries:
the University of Hasselt and Maastricht University.

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Testing non-autonomous models of antibody dynamics by parametric fitting of data on HAV vaccination: Exploratory study

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. OLIVIER LEJEUNE

Abhishek Bakuli

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Testing non-autonomous models of antibody dynamics by parametric fitting of data on HAV vaccination: Exploratory study

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. OLIVIER LEJEUNE

Abhishek Bakuli

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Abstract

The immunological system existing in our body can be considered as a dynamical system, comprising of antibodies, and plasma cells. It starts with the development of antibodies, which finally evolve into plasma cells, which can be long lived memory cells providing long term immunity or short lived plasma cells providing short term immunity. The antibody count in our blood can be observed, and hence this entire dynamical system can be considered as analogous to the population dynamical system in demography. There is an antibody population which decreases with time, and there is development of plasma cells, which increase in number initially but decrease with age. This is similar to a growth and decay process in demographical models. The project is concerned with studying the growth and decay process existing within the immunological system mathematically. Differential equations are popular in the study of dynamical systems, and with the observed empirical data of the antibody count post vaccination of Hepatitis A, in particular for this project; we want to look at statistical methods that exist, in estimating the parameters of the differential equation system. We look at two broad cases, one when the system parameters governing growth and decay are age independent, and one where they are age dependent.

In this project, we have presented an overview of the various statistical methodologies that exist in looking at such problems. We also present the problems that arise, when cases like analytically closed functional forms do not exist for the solutions of the system of differential equations. We have looked into several optimization strategies, in particular the Stochastic Approximation of the EM algorithm method, in order to estimate the parameters from a non linear mixed effect model. This method is not a gold standard, and thus has its share of problems which could be solved with further research. We could obtain the parameter estimates; however we had difficulties in obtaining the estimates of the standard errors, through the Fisher Information Matrix. However visual predictive tools have been used to explore if the parameter estimates were fitting the data well or not. Numerically we obtain positive values for the age dependent antibody kinetic parameters, which indicate that the production rate of plasma cells decrease with age and the corresponding decay rate increases with age; however the standard errors could not be computed to establish the statistical significance of these estimates. There also exists a problem related to having standard software/program/package developed for handling such problems in general, which could be a topic of research interest in the future.

Acknowledgement

I would like to thank all my parents, relatives, teachers and friends, who have been a big support for me all through my life and helped me in transforming me to the person I am today. I would specially like to thank Prof. Dr. Niel Hens for his amazing support and encouragement all along the thesis period, and Prof. Dr. Olivier Lejeune also for his remarkable guidance on the essential explanations of the differential equation system and relating it to the biological phenomena, for the mathematical component of the project. Also to mention the models that have been used in the project, are a part of his research work in this topic and the reference paper from PloS Computational Biology, 2012 on Antibody Dynamics, was co authored by both of them among others, and I was able to learn a lot from both my supervisors.

I would also specially like to thank Prof. Dr. Bhimashankaram Pochiraju, and Prof. Dr. Pulak Ghosh, who have been a big inspiration in my life for pursuing Applied Statistics, in particular Biostatistics, as a study discipline. Also I would like to thank Prof. Dr. Christel Faes, who introduced to me to this program when I was in India, and the Vlaamse Interuniversitare Raad(VLIR) university cooperation development program for financially supporting me for the last two years of my study program in Belgium.

Finally I would like to thank CenStat, Universiteit Hasselt for all the valuable knowledge that I have gained as a part of this program. Special mention to all my group mates, from Statistics as well as Transportation Sciences, in particular Eleni Kourkouni and Joram Langbroek, who were there with me through thick and thin over the two years, and made learning statistics, a joyful experience.

Abhishek Bakuli
September 2013
Genk, Belgium

Contents

Introduction

Immunity is largely understood as our resistance against foreign antigens; infection, disease or any unwanted pathogen. The key feature of mammalian defense and maintenance is adaptive immunity: body's capability to modify specifically towards each attack. Following vaccination or infection, T cells of the immune system activate the B-lymphocytes for the specific antigen and glycoprotein's that recognize the antigen and tag it for elimination i.e. antibodies (Ab) are released. Ab are also called immunoglobulin and occur in distinct idiotypes of G, A, M, E and D.

What makes this system remarkable is its ability to remember the identity of a pathogen. Inactive Memory B cells do not actively secrete Ab but instead maintain their immunoglobulin in the membrane-bound form that serves as the antigen-specific B-cell receptor. Whereas, plasma cells are differentiated cells that no longer express surface-bound Ab but continuously secrete them without requiring further antigenic stimulation (Ahmad and Gray, 1996). Upon activation, the secondary response thus elicited is different than the primary response at 3 major aspects: (i) rapidity, (ii) relatively more immunoglobulin G (IgG), IgA, or IgE than IgM, and (iii) higher affinity (Amanna et. al, 2010).

Hepatitis A is an acute infection of the liver, caused by the Hepatitis A virus (HAV) and is usually transmitted via fecal-oral route. Since 1995, hepatitis A vaccines have been used to prevent hepatitis A in people not exposed to the hepatitis A virus previously. Only three of the included trials were considered to be at low risk of bias; free from overestimation of benefits and underestimation of harm due to systemic errors. In persons not previously exposed to hepatitis A infection, hepatitis A vaccination with inactivated or live attenuated hepatitis A vaccines had a clear effect on reducing the risk of developing clinically apparent hepatitis (Irwing et. al, 2012).

Antibodies that persist following vaccination have long been considered as the principle marker of protection against hepatitis A infection. In extensive studies of children and adults, the inactivated hepatitis A vaccine has been found to be highly immunogenic but its safety remains to be an issue of constant controversy (Irwing et. al, 2012 & Demichili and Tiberti, 2003).

The field of systems biology is an ever emerging approach in research based on biological sciences. It started with the onset of mathematical models for population studies, and this has a major influence in studying the immunological system of the body. In an immunological system, we study the antibody count in our blood, which protects our body from a variety of diseases. The decrease in the count of antibodies, can severely affect our health condition, and thus population studies at the level of antibodies is a prime area of research. A lot of complex biological phenomena, like the antibody system in our body can be looked upon as a dynamical system mathematically, which is expressed in terms of systems of partial or ordinary differential equations. With the advent of modern technology, we also have access to observe a lot of biological phenomena empirically, and statistical methodology helps us to understand the dynamical systems better with observed empirical evidence.

The statistical analysis of the dynamical systems governing the functioning of the immunological systems specially the antibody dynamics is a relatively new research field. The work dating back to 2007 by Park et. al (2007), in the Journal of Royal Society Interface, has shown a statistical study for the dynamics of the antibody loss due to a disease causing agent. There has also been works by Andraud, Lejeune et. al (2012), where they have described the life span of antibodies, short lived plasma cells and long lived plasma cells in describing the observed antibody kinetics after Hepatitis A Vaccination (HAV) which is the main system of study in this project. This work was based on certain assumptions in relation to the decay rate of antibodies, and here we look at exploring the possibilities of relaxing assumptions and looking at complexities that develop in this process.

We present our work as a discussion on existing statistical methods and its implementation in common software packages, and what are the problems that are incurred when assumptions in the antibody dynamical system are changed. We also pose certain unsolved problems that could be looked into for further research.

Theory and Methods

Functional Data Analysis

Functional data come in many forms, but their defining quality is that they consist of functions—often, but not always, smooth curves. Functional data arise in many different fields, ranging from the shapes of bones excavated by archaeologists, to economic data collected over many years, to the drug reaction mechanism in Pharmacokinetics as well as in modeling growth in different time scales. The fundamental aims of the analysis of functional data are the same as those of more conventional statistics: to formulate the problem at hand in a way amenable to statistical thinking and analysis; to develop ways of presenting the data that highlight interesting and important features; to investigate variability as well as mean characteristics; to build models for the data observed, including those that allow for dependence of one observation or variable on another, and so on (Ramsay and Silverman , 2002).

Functional data arise as the original observations are interpolated from longitudinal data, quantities observed as they evolve through time, especially in the example that we are considering in our study. However, there are several other ways that can result in development of functional data. Instead of looking at data individually, in Functional Data Analysis (FDA), the data set is converted into a set of infinite-dimensional curves. There are several kinds of complications that arise in functional data, for example in our case; we do not have observations that are equally spaced and we only observe at discrete time points; however it is a desirable property to use functions to reflect smooth variations in the measured variable. There is a major use of derivatives in this field of analysis, and several data features can be better illustrated through the derivatives of a specific order, since we try to think of the records as functional observations as compared to observations measured in discrete time (Wang, 2007 & Ramsay and Silverman, 2005).

In FDA, we have noisy observations on discrete time points, and it can be represented by a linear combination of basis functions. Commonly used basis systems are splines, Fourier basis, and polynomial basis, just to name a few. However basis approximation for the functional data is good, under the condition that the basis functions can describe the essential characteristics of the observed data.

Smoothing is the process of converting discrete observations into functions, and it is assumed that discrete values are subject to observational error. The method of Roughness Penalized Smoothing is a commonly used technique to approximate functional data and control for the smoothness along with estimating the derivatives. After the smoothing process, we reduce the dimension from, the number of observations per subject say n, to the number of basis functions that are used for the approximation of the data say m. Now if m is larger than n, because of the problems in approximating the underlying functions, due to the sharp changes or discontinuity in the data, a penalty term is often used to control the roughness of the functions and in turn avoid the problem of over fitting. Often differential equations are used to define the penalty term in penalized smoothing, leading to better estimates for smooth functions and their derivatives (Ramsay and Silverman 2002 & Wang, 2007).

In contrast to Time Series Analysis, where it is more common to have equally spaced observations and differencing is a widely used concept, in FDA, there is more common usage of the derivatives (Ullah and Finch, 2013). In comparison to Longitudinal Data Analysis (LDA), FDA requires more frequent observations. " FDA tend to be exploratory in order to represent and display data in order to highlight interesting characteristics, perhaps as input for further analysis − whereas those of LDA have a stronger inferential component. This contrast can be seen in the use of estimated time correlation functions in the two areas; correlation functions are used in the FDA literature in a descriptive manner to characterize time dependencies in the curves, whereas an important aim in the LDA literature of estimating these correlation functions is to draw valid inferences." , (Rice, 2004).

However there are several similarities in LDA and FDA also, like the characterization of the marginal profile over time, estimations of the individual profiles, checking for variability patterns in the curves, and assessing the relationship of the shape of curves to the covariates (Rice, 2004).

Non Linear Mixed Effect Models

When we have data, for a continuous response that evolves over time, within individuals from a population of interest, we have the framework of repeated measures or longitudinal data. Mixed effect models for repeated measures data have become a popular analysis tool, because it has a flexible covariance structure that allows for non constant correlation among observations, and also works for unbalanced data. The intuitive appeal of mixed effect model is because, we assume that individual responses follow a common functional form, and parameters vary among different individuals, which is plausible in several scenarios of life.

Much of the research in this area is based on the field of linear mixed effect models, where the fixed effect terms are a linear function of the parameters of the model.

$$\beta_0 + \beta_1 x_{1j} + \ \dots + \beta_p x_{pj},$$

where $x_i^{'}s$ are the observed covariates for the j^{th} individual, and β_i's the parameters.

However in several situations like dose response modeling, or pharmacokinetics or growth models and many more situations, often require non linear functions in the parameters. Non Linear Mixed Effect (NLME) models are the generalization of the linear mixed effect model framework, as well as the non linear model framework.

Let Y_{ij} denote the j^{th} observed response for the i^{th} individual measurement at time point t_{ij} , i = 1, 2, · · · ,N, j = 1, 2, . . . ,n_i.
In linear mixed models, the mean is modeled as a linear function of regression parameters and random effects

$$E(Y_{ij}|b_i) \ = \ x_{ij}'\beta \ + \ z_{ij}'b_i$$

In generalized linear mixed models, apart from a link function, the mean is again modeled as a linear function of regression parameters and random effects

$$E(Y_{ij}|b_i) \ = \ h(x_{ij}'\beta \ + \ z_{ij}'b_i)$$

For the NMLE models, they are no longer modeled as a function of the linear predictor $x_{ij}'\beta \ + \ z_{ij}'b_i$. In this project, we assume that the conditional distribution of Y_{ij}, given b_i is belongs to the exponential family. The mean is modeled as

$$E(Y_{ij}|b_i) \ = \ h(x_{ij}, \beta, z_{ij}, b_i)$$

The random effects are assumed to be normally distributed with mean 0 and variance D. $fij(y_{ij}|b_i, \beta, \varphi)$ is the conditional density of Y_{ij} given b_i , and $f(b_i|D)$ be the density function of the N(0,D) distribution. In most cases of NLME models the obtained likelihood does not have a closed form solution, and hence the maximum likelihood cannot be obtained analytically to obtain the parameter estimates. Hence we need numerical algorithms to solve this problem (Verbeke and Molenberghs, 2010 & Lindstorm and Bates, 1990).

In the literature, many methods have been studied to solve this problem for the integration of the likelihood, and one of the most common methods is linearization of the non linear model. Generally, these methods differ in their assumptions regarding the distribution of the random effects, including the inter- and intra-individual variability, and in approximations used to deal with inter-individual random effects. Several methods like first order method, conditional first order linearization, Laplacian approximation, Lindstorm and Bates algorithm and many others, approximate the non linear function to a linear function. A lot of these methods were developed for solving the problems in population pharmacokinetics.

An alternate approach to solving such problems in the Bayesian perspective has also been explored in the literature. The NLME model is formulated as a three stage Bayesian hierarchical model.

$$\begin{aligned} Y_i|x_i, b_i &\sim N(g_i(x_i, \beta_i), H_i(\beta_i, x_i, \theta)) \\ b_i|\beta, \Sigma &\sim N(0, \Sigma) \\ (\beta, \Sigma, \theta) &\sim \pi(\cdot) \end{aligned}$$

where $\beta_i = A_i\beta + B_ib_i$ or more generally $\beta_i = h(a_i, \beta, b_i)$. We assume that the functional forms of g_i and H_i are analytically known as a function of parameters. Inference is drawn by computing the posterior distribution of the parameters given the observed data assuming hyperprior distributions of the parameters. This method is based on the calculation of the full conditional probability distribution of the parameters (Ghosh, 2010 & Wang, 2007). As always in a Bayesian framework, knowledge from previous studies can be utilized, by specifying appropriate prior distributions for the parameters. Here all the parameters, both fixed and random effect parameters are assumed to be random, since they are assumed to have probability distributions. The Bayesian approach can also handle missing and observations quite easily.

Within the framework of NLME, we are often interested in describing the mean profile (or trajectory) as a dynamic relationship between response and an explanatory variable (for example time), by a system of ordinary differential equations (ODE) whose parameters describe the different characteristics of the underlying population. ODE's are commonly used tools to describe a dynamical system, where we are interested in modeling the rate of change over time than, static average value of the response variable. In reality it turns out that rarely are we able to derive a closed form expression for the exact solution for such a system. The absence of a closed form analytical solution for the system of ODE's makes parameter estimation in such models challenging and computationally demanding (Goyal and Ghosh, 2006).

There are many methods that have been proposed in literature to deal with this problem of estimating parameters in ODE's without closed form solutions; however they suffer from several drawbacks when fitting to noisy data. The most commonly used method is the non linear optimization procedure. This is a very computationally intensive procedure, since the ODE's have to be solved numerically in a repeated manner when updating the parameter values and the initial values of the components. Often the initial values of the components are not known and add to the list of parameters to be estimated. The process also depends heavily on the quality of the starting values for the parameters and components. Sometimes algorithms are trapped in local minima and thus the ODE's become unsolvable.

One possible approach to counter the above problem is through the Bayesian methodology, which uses the integrated Euler approximation, to obtain a closed form approximation of the mean function, without imposing restrictive conditions on the system of ODE's. The major advantage in this method is that, there is no need to keep evaluating the numerical approximate solution for the mean function, repeatedly for interpolation or extrapolation. This technique is called the Bayesian Euler Approximation Method (BEAM) (Goyal and Ghosh, 2006). However this method is also often computationally intensive and it is also difficult to select initial values and appropriate priors in advance.

The likelihood approximations often perform well if the number of intra individual observations is not small and the variability of the random effects is not large. Deviations from the above case might result in considerable amount of errors (Davidian and Giltinan, 1995 & Pinheiro and Bates, 1995 & Lindstrom and Bates, 1990). Thus the use of exact methods like the Monte Carlo methods for likelihood based estimation came into existence. The EM algorithm (Rubin et. al, 1977) provides a tool for obtaining maximum likelihood estimates, under models that yield formidable likelihood equation. The EM algorithm is an iterative routine requiring two primary calculations each iteration; Computation of a particular conditional expectation of the log-likelihood (E-step) and maximization of this expectation over the relevant parameters (M-step). The Monte Carlo EM (MCEM), introduced by Wei and Tanner, 1990, is a modification of the EM algorithm where the expectation in the E-step is computed numerically through Monte Carlo simulations. Although the Monte Carlo estimate presents a tractable solution to problems where the E-step is not available in closed form, we must account for the additional Monte Carlo (MC) error inherent in the approach and try to minimize the increased computational cost in obtaining the MC sample (Levine and Casella, 2001). In the MCEM process for the E step approximation is done using the simulated samples from the exact conditional distribution of the random effects given the observed data, and this is a popular estimation procedure in mixed models. Walker in 1996 proposed MCEM algorithm using approximations based on samples from the distribution of the random effects for ML exact estimation in a specific class of NLME models. A stochastic version of the EM algorithm (SAEM) using stochastic approximations involving samples obtained via Markov chains for fitting NLME models is proposed by Kuhn and Lavielle in 2005. The Fischer Information matrix can also approximated stochastically by this method. The biggest advantage of this method is that it converges to the neighborhood of the maximum likelihood estimate very quickly over the other methods, and thus the confidence interval of this estimate can be obtained very quickly. The SAEM can be used for both homoscedastic as well as heteroscedastic models. In heteroscedastic models, the estimates for the fixed effect models are estimated in a Bayesian framework in terms of their expectations (Wang, 2006). Over all the SAEM procedure gives us estimates close to the Maximum Likelihood estimates in very little iterations. The hypothesis testing is done by using the Wald Test.

The other alternative method is to look at Inverse Problems. It is a general methodology to convert observed measurements into information about a specific system of interest. We can formulate this in the following way, where we want to find the best "*m*" such that

$$d = G(m),$$

where *G* is an operator describing the explicit relationship between the observed data, *d* and the model parameters. When we have a discrete linear inverse problem describing a linear system, *d* (model parameters) and *m* (the best model) are vectors, and *G* is the observation matrix. We can realize the simple linear regression model from the above description. Thus we can finally obtain

$$m = G^{-1}\,(d)$$

However in practical scenarios, very often G is non invertible. Thus we have to resort to optimization methods to solve the inverse problem. We generally define a target known as the objective function for the inverse problem. We objectively try to measure the difference between the observed and the fitted value, and try to minimize it at the end. The objective function at the end becomes

$$\varphi = ||d - G(m)||_2^2$$

In the Linear inverse problem, m is the ordinary least squares estimate. The non linear inverse problem is not that straight forward, and the solutions depend on many other conditions. The major purpose although is to find the global minima. However there exist many challenges, such as, many models comprise non-identifiable parameters which cannot be unambiguously determined with sufficient precision (Vajda, Rabitz, Walter, and Lecourtier, 1989). Such non-identifiability is manifested by functionally related parameters, such that the effect of altering one parameter can be, at least partly, undone by altering some other parameter(s). This type of over parameterization is common for complex models and especially in biological modeling; it is nearly unavoidable (Mieleitner and Reichert, 2006). In order for the data fitting algorithms to converge, and for the parameters to be estimated with reasonable precision, the parameter set must be identifiable (Soetaert and Petzoldt, 2010).

Models of Antibody Dynamics

In this section we look at the system of Immunological Process, mathematically, in our body post vaccination, through a set of differential equations whose solutions result in functions which are non- linear in the parameters. We also know that immunity is an evolving process with time and external sources like vaccination. Antibodies are produced in our body, as well as their population decays with age too. This decay is the result of observing decreased immunity with age. Vaccines also help us develop immunity, some of which stay for the entire life time, through the process of being transformed into different kind of plasma cells. This entire process is mathematically described below.

Age-independent model (by Andraud , Lejeune, et. al, 2012)

Here we look at models, where the kinetic parameters are not dependent on age.

The evolution equation for the population of long-lived plasma cells P_l is

$$\frac{dP_l}{dt} = -\mu_l P_l ,$$

where μ_l represents their average decay rate. Integrating the equation gives

$$\int \frac{dP_l}{P_l} = -\int \mu_l dt \quad \ln P_l = -\mu_l t + c \quad P_l(t) = P_l(0)e^{-\mu_l t} .$$

The solution for the population of short-lived plasma cells P_s is similarly

$$\frac{dP_s}{dt} = -\mu_s P_s \quad \Rightarrow \quad P_s(t) = P_s(0)e^{-\mu_s t} ,$$

where μ_s represents their average decay rate.

The evolution equation for the population of antibodies A is

$$\frac{dA}{dt} = F(t) - \mu_a A \quad F(t) = \varphi_l P_l(t) + \varphi_s P_s(t) ,$$

where μ_a represents their average decay rate; φ_l and φ_s are the production rates of antibodies by long- and short-lived plasma cells respectively.

The solution of the homogeneous differential equation, corresponding to $F(t)=0$, is

$$\frac{dA}{dt}=-\mu_a A \quad \Rightarrow \quad A(t)=A(0)e^{-\mu_a t}\ .$$

Applying the method of variation of constants, the solution of the inhomogeneous differential equation is looked in the form

$$A(t)=K(t)e^{-\mu_a t}\ .$$

By substitution it follows that

$$\frac{dK}{dt}e^{-\mu_a t}-\mu_a K(t)e^{-\mu_a t}=F(t)-\mu_a A(t)\,.$$

After simplification one gets

$$\frac{dK}{dt}=F(t)e^{\mu_a t} \quad \Rightarrow \quad K(t)=K(0)+\int_0^t F(z)e^{\mu_a z}dz\,.$$

Hence the solution of the inhomogeneous differential equation is

$$\begin{aligned} A(t) \quad &= \left[A(0)+\int_0^t \left(\varphi_l P_l(z)+\varphi_s P_s(z)\right)e^{\mu_a z}dz\right]e^{-\mu_a t} \\ &= \left[A(0)+\varphi_l\int_0^t P_l(z)e^{\mu_a z}dz+\varphi_s\int_0^t P_s(z)e^{\mu_a z}dz\right]e^{-\mu_a t}\,. \end{aligned}$$

Substituting the solutions of the plasma cells populations it becomes

$$A(t)=\left[A(0)+\phi_l\int_0^t e^{(\mu_a-\mu_l)z}dz+\phi_s\int_0^t e^{(\mu_a-\mu_s)z}dz\right]e^{-\mu_a t},$$

where we have set $\phi_l=\varphi_l P_l(0)$ and $\phi_s=\varphi_s P_s(0)$.

Based on the above equation we come up with three models which are as follows

Model 1: Complete Model

$$A(t)=\frac{\Phi_s}{\mu_a-\mu_s}\,e^{-\mu_s t}+\frac{\Phi_l}{\mu_a-\mu_l}\,e^{-\mu_l t}+\left(A_0-\frac{\Phi_s}{\mu_a-\mu_s}-\frac{\Phi_l}{\mu_a-\mu_l}\right)e^{-\mu_a t}$$

Model 2: Asymptotic Model, assuming life span of long lived plasma cells as infinity, or $\mu_l=0$

$$A(t) = \frac{\Phi_s}{\mu_a - \mu_s}\, e^{-\mu_s t} + \frac{\Phi_l}{\mu_a} + \left(A_0 - \frac{\Phi_s}{\mu_a - \mu_s} - \frac{\Phi_l}{\mu_a}\right) e^{-\mu_a t}$$

Model 3: Plasma Cell Driven Kinetic Model, assuming that the antibody life span is short relative to the plasma cell life span ($\frac{\mu_{l,s}}{\mu_a} \ll 1$)

$$A(t) = \beta_s e^{-\mu_s t} + \beta_l e^{-\mu_l t}$$

Where $\beta_s = \frac{\Phi_s}{\mu_a}$ and $\beta_l = \frac{\Phi_l}{\mu_a}$.

Age- dependent Models (Hoppenstead, 1997)

Unlike the previous class of models, here we assume that the time dependence of the kinetic parameters appears only through age, that is to say time since generation, τ. In other words, we consider the internal decay processes of the biological agents (plasma cells and antibodies) in an otherwise stationary physiological state with respect to the specific humoral immune response.

The evolution equations for the populations of long- and short-lived plasma cells are now

$$\begin{array}{ll}
\frac{\partial P_{l,s}}{\partial t} + \frac{\partial P_{l,s}}{\partial \tau} = -\mu_{l,s}(\tau) P_{l,s}(\tau,t), & \forall \tau \geq 0, \forall t > 0; \\
P_{l,s}(0,t) = G_{l,s}(t) = 0, & \forall t > 0; \\
P_{l,s}(\tau,0) = P_{l,s_0}(\tau), & \forall \tau \geq 0.
\end{array}$$

We assume that the generation of new plasma cells is ended, $G_{l,s}(t) = 0$.

Hence the solutions are

$$P_{l,s}(\tau,t) = \left\{\begin{array}{ll}
P_{l,s_0}(\tau - t) e^{-\int_{\tau-t}^{\tau} \mu_{l,s}(z)dz} & \forall \tau \geq t \\
G_{l,s}(t-\tau) e^{-\int_0^{\tau} \mu_{l,s}(z)dz} = 0 & \forall t > \tau
\end{array}\right\}.$$

The evolution equation for the population of antibodies is now

$$\begin{array}{ll}
\frac{\partial A}{\partial t} + \frac{\partial A}{\partial \tau} = -\mu_a(\tau) A(\tau,t), & \forall \tau \geq 0, \forall t > 0; \\
A(0,t) = G_a(t), & \forall t > 0; \\
A(\tau,0) = A_0(\tau), & \forall \tau \geq 0.
\end{array}$$

The production of new antibodies is given by the integral

$$
\begin{aligned}
G_a(t) &= \int_0^{+\infty} \left[\varphi_l(\tau)P_l(\tau,t) + \varphi_s(\tau)P_s(\tau,t)\right]d\tau \\
&= \int_t^{+\infty} \varphi_l(\tau)P_{l_0}(\tau-t)e^{-\int_{\tau-t}^{\tau} \mu_l(z)dz} d\tau + \int_t^{+\infty} \varphi_s(\tau)P_{s_0}(\tau-t)e^{-\int_{\tau-t}^{\tau} \mu_s(z)dz} d\tau
\end{aligned}
$$

The solution is

$$
A(\tau,t) = \begin{Bmatrix} A_0(\tau-t)e^{-\int_{\tau-t}^{\tau} \mu_a(z)dz} & \forall\tau \geq t \\ G_a(t-\tau)e^{-\int_0^{\tau} \mu_a(z)dz} = 0 & \forall t > \tau \end{Bmatrix}
$$

The observable variable is the total number of antibodies irrespective of their time since production, that is to say the integral

$$
\overline{A}(t) = \int_0^{+\infty} A(\tau,t)d\tau = \int_0^{t} G_a(t-\tau)e^{-\int_0^{\tau} \mu_a(z)dz} d\tau + \int_t^{+\infty} A_0(\tau-t)e^{-\int_{\tau-t}^{\tau} \mu_a(z)dz} d\tau
$$

We get the integral expression of the observable antibody count by substitution and is of the form as expressed below.

$$
\overline{A}(t) = e^{-\mu_a t}\left[\int_0^t \left(\overline{P}_l(0)\varphi_l(z)e^{-\int_0^{z} \mu_l(z')dz'} + \overline{P}_s(0)\varphi_s(z)e^{-\int_0^{z} \mu_s(z')dz'} \right) e^{\mu_a z}dz + \overline{A}(0)\right].
$$

where,

$$
\overline{A}(0) = \int_0^{+\infty} A(\tau,0)d\tau = \int_0^{+\infty} A_0(\tau)d\tau
$$

Let us assume that both the increases of the plasma cells decay rates and the decrease of the antibody production rates are exponential. If we set

$$
\mu_{l,s}(\tau) = \mu_{l,s}(0)e^{\eta_{l,s}\tau} \quad \text{and} \quad \varphi_{l,s}(\tau) = \varphi_{l,s}(0)e^{-\gamma_{l,s}\tau},
$$

The asymptotic model corresponds to a static population of long-lived plasma cells with infinite lifespan and steady antibody production rate. If we set

$$\mu_l(\tau) = 0 \quad \text{and} \quad \varphi_l(\tau) = \varphi_l,$$

We obtain the asymptotic model as follows.

$$\overline{A}(t) \; = \frac{\phi_l}{\mu_a} + e^{-\mu_a t}\left[\Theta_s \int_0^t e^{(\mu_a - \gamma_s)z - \rho_s e^{\eta_s z}} dz + \overline{A}(0) - \frac{\phi_l}{\mu_a}\right]$$

In the above equation the age dependent kinetic parameters are γ and η.

In case of the age dependent kinetic parameters, we have a closed form analytical solution for the Plasma Cell Driven Kinetic model. Here we also assume $\frac{\varphi_{l,s}}{\mu_a}$ is a finite quantity. The two models are based on the assumptions of the functional form for the plasma cell production and decay.

The general form is

$$A(t) = \frac{P_l(0)\varphi_l(t)}{\mu_a} \; e^{-\int_0^t \mu_l(z)dz} + \frac{P_s(0)\varphi_s(t)}{\mu_a} \; e^{-\int_0^t \mu_s(z)dz}$$

Model 1: $\varphi_{l,s}(\tau) = \varphi_{l,s}(0)e^{-\gamma_{l,s}\tau}$ and $\mu_{l,s}(\tau) = \mu_{l,s}(0)e^{\eta_{l,s}\tau}$, for simplicity $\gamma_{l,s} = 0$, which implies a constant rate of production. Exponential function for the decay.

$$A(t) = b_l \; \exp\left(\frac{-\mu_l(0)(e^{\eta_l t})}{\eta_l}\right) + \; b_s \; \exp\left(\frac{-\mu_s(0)(e^{\eta_s t})}{\eta_s}\right)$$

where $b_{l,s} = \frac{P_{l,s}(0)\varphi_{l,s}(0)}{\mu_a}$. $\eta_{l,s}$ are the age dependent kinetic parameters we are interested in.

Model 2: $\varphi_{l,s}(\tau) = \varphi_{l,s}(0)e^{-\gamma_{l,s}\tau}$ and $\mu_{l,s}(\tau) = \mu_{l,s}(0) + \beta_{l,s}\tau^{\eta_{l,s}}$, for simplicity we assume $\gamma_{l,s} = 0$, which implies a constant rate of production. Polynomial function for the decay.

$$A(t) = b_l \; \exp\left(-\,(\mu_l(0)t + \; \beta_l \frac{t^{\eta_l+1}}{\eta_l + 1})\right) + \; b_s \; \exp\left(-\,(\mu_s(0)t + \; \beta_s \frac{t^{\eta_s+1}}{\eta_s + 1})\right)$$

where $b_{l,s} = \frac{P_{l,s}(0)\varphi_{l,s}(0)}{\mu_a}$. $\eta_{l,s}$ are the age dependent kinetic parameters we are interested in.

Model 3: We consider the case of Model 1, when $\gamma_{l,s} \neq \; 0$

$$A(t) = b_l \; \exp -\left(\gamma_l t + \frac{\mu_l(0)(e^{\eta_l t})}{\eta_l}\right) + \; b_s \; \exp -\left(\gamma_s t + \frac{-\mu_s(0)(e^{\eta_s t})}{\eta_s}\right)$$

$\eta_{l,s}$ and $\gamma_{l,s}$ are the age dependent kinetic parameters we are interested in.

Model 4: We consider the case of Model 2, when $\gamma_{l,s} \neq 0$

$$A(t) = \frac{b_l}{1+\alpha_l t^{\gamma_l}} \exp\left(-\left(\mu_l(0)t + \beta_l \frac{t^{\eta_l+1}}{\eta_l+1}\right)\right) \quad + \quad \frac{b_s}{1+\alpha_s t^{\gamma_s}} \exp\left(-\left(\mu_s(0)t + \beta_s \frac{t^{\eta_s+1}}{\eta_s+1}\right)\right)$$

$\eta_{l,s}$ and $\gamma_{l,s}$ are the age dependent kinetic parameters we are interested in.

Statistical Software / Packages for FDA

Here is a description of some of the software that exists for dealing with problems of Functional Data Analysis as a broad field of study. We shall look at R and SAS only for this report.

The *nlme* package in R is a standard tool for analyzing non linear mixed effect models in R. It was developed by Pinhero and Bates in 1990. This package however cannot deal with functions that do not have an analytical closed form. The other drawback of this package is that it cannot deal with systems of differential equations.

The *nlmeODE* package developed by Tonroe (2004), enables *nlme* package to handle the system of ODE's. The estimation of the parameters is done by solving the ODEs numerically for every iteration in the numerical estimation process of the mixed effects model, which can take a very long time, even on fast computers (Hagenbuch, 2011). This package is also tailor-made for specific type of problems, particularly in pharmacokinetics, and requires a very specific way the data set needs to be structured and labeled. The error displayed was that "object of type 'closure' is not subsettable", although there was no error in defining the functional form, which is one of the most common instances where this error occurs. The same function was used to solve the ODE system at hand, and it worked perfectly. The other errors that were common to both this and *nlme*, was that "Step halving factor reduced below minimum in PNLS step", and "Maximum number of iterations reached without convergence".

Within R, there exists a package *deSolve* (Soetaert, et. al, 2010) to solve a system of differential equations. The dynamical system consisting of ODE's can be solved by this program, by entering starting values for the system. However only three specific kinds of PDE's can be solved in this package, and there is no special provision for solving a system of linear PDE as we have here.

Further development in terms of computational efficiency to solve a non analytical likelihood function which is common in a non linear mixed effect model, a package *saemix* (Comets, et. al, 2013) uses the stochastic approximation EM algorithm, for parameter estimation. This is sufficiently faster than most of the other existing methods for parameter estimation procedures, since the population parameter estimation does not require approximation of the likelihood function by the SAEM algorithm and the dependence on the initial starting values is much less. This package also requires closed form analytical solutions of the objective function of interest. Having too many parameters to estimate may result in failure of the estimation procedure though. The model fitted with the Gaussian Adaptive Quadrature is much more time consuming than with the importance sampling method, and the results are quite close to each other in terms of parameter estimates and AIC. However at least for the examples we verified, the AIC under Importance sampling was always less than that of Gaussian Quadrature method.

The *FME* package (Soetaert and Petzoldt, 2012) is another improvement in the field of functional data analysis, where it uses the method of inverse modeling, to fit the data to the function, and it works even when closed form analytical solutions to the system of equations do not exist. This package has the drawback that it cannot handle longitudinal data. Also, since it is related to the deSolve package, by virtue of the authors of both packages being the same people, it has no specific tool for handling linear system of PDE's.

For solving the function expressed as an integral equation, as in case of the asymptotic model, that has been discussed, no specific package exists as of now to handle such problems. One way to approach the problem is to perform the optimization of the completely specified likelihood function or use non linear least squares on the user defined function. We tried to optimize the user defined likelihood function, assuming the Gaussian distribution, as well as define a function (for the non linear least squares), to obtain the estimates, however there were issues with the starting values, or the function was non convergent.

SAS, also has the PROC NLMIXED, which handles the non linear mixed effect models, with a closed analytical functional form, and gives us estimates to the population parameters, as well as the subject specific parameters using the numerical integration approximation method, Gaussian Adaptive Qaudrature as the default method to optimize the objective function. This works in an iterative process, finding the maximum likelihood estimates of the fixed effects for these values, re-estimating the random effects, and then going back to the fixed effects. Thus it is a very time consuming process, and it is also highly dependent on the starting values (Pillai, et. al, 2005).

"Summary of the different programs/packages used"

R	**Nlme**	linearization based algorithm for optimization, slow convergence, needs analytically closed functional forms, can handle hierarchical data, cannot handle Differential Equations directly
	nlmeODE	combines aspects of nlme and odesolve package, works only for ODE's, analytically closed functional forms not needed, often problems are incurred in functional specification, and not very popular
	saemix	uses Gaussian Adaptive Qaudrature, as well as Stochastic Approximation of EM algorithm for optimization, can handle only mixed effect models, one of the fastest optimization methods, can handle only analytically closed functional forms
	FME	uses the concept of Inverse Modeling, does not handle hierarchical data-in fact only one observation of a variable at a specific time is considered, handles wide variety of dynamical systems but only three specific kinds of PDE's
SAS	**NLMIXED**	uses Gaussian Adaptive Quadrature method for optimization, quite reliable and most popularly used tool for mixed effect models, needs analytically closed functional forms, convergence is often very slow

However in general when working with non linear models (even mixed effect models included) in both R and SAS, there is dependency on the starting values of the parameter estimates. Often it is seen that the algorithms, get stuck in the local minima, and hence has issues with convergence. We also did not explore the software implementing the Bayesian Methodology for this report.

Dataset Description

We have dataset from a longitudinal study, of healthy HAV-seronegative adults aged between 18 and 40 years, and they were enrolled after giving a written informed consent. We have 284 patients who have been administered the Harvix^{TM} 1440, which is a hepatitis A vaccine. We have a longitudinal profile of these patients; with the outcome being measured is the population of antibodies present in the blood, with the unit of measurement as mIU/ml. We are interested in the long term persistence of antibodies, after full vaccination schedule, and observations are taken at 1,12,18,24,36,42,48,50,66,78,90,102,114 and 126 months after boosting. We do not have any other covariate of interest in the data set. This data set is the observed empirical evidence that we have about the antibody system existing in our body.

Exploratory Data Analysis

We look at the histogram of the antibody count (population), and observe that it is highly skewed, where as the log base 10 transformed counts has a symmetric distribution.

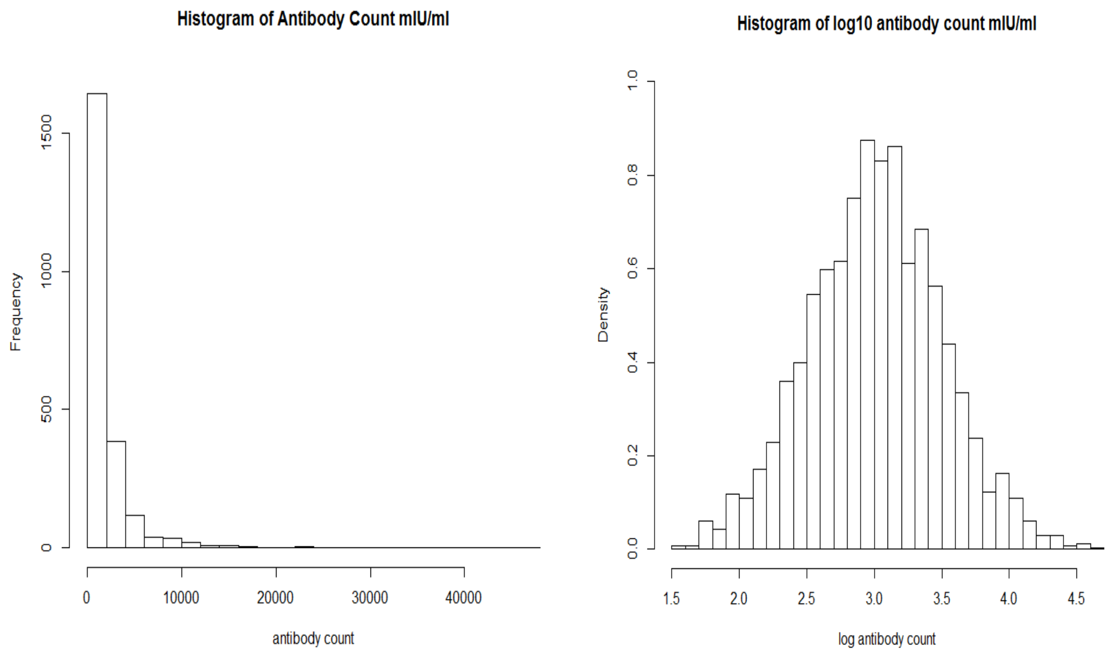

Figure 1:Histogram of the observed Antibody count(left), log$_{10}$Antibody count(right)

Thus we have a rough idea that the antibody count can be thought of as log normally distributed, or the log scaled antibody count is normally distributed. However the residuals in the statistical models are not necessarily of the same distributional form as that of the observed data.

We also present the table of the number of observations at every time point, and also the average antibody count at each of the time points. We see that there is a sharp drop from time point 1 to time point 2, and in the following interval till time point 3, after which the change is quite minimal. This can also be seen in the graphical representation of the profiles, thereby indicating that a non linear function would be required to fit the data.

Time	Average(Log$_{10}$Y)	Average(Y)	Number of Observations
1	3.631	6468.122	262
12	3.173	2231.772	237
18	3.059	1565.400	70
24	3.057	1839.340	147
30	2.898	1159.635	74
36	3.022	1717.252	143
42	2.783	973.182	66
48	2.903	1379.503	143
50	2.755	868.647	68
66	2.836	1141.693	192
78	2.897	1222.093	204
90	2.799	1022.508	184
102	2.903	1253.359	170
114	2.904	1248.195	159
126	2.847	1020.860	157

Table 1: Summary of the Population Averaged, observed data at the different time points.

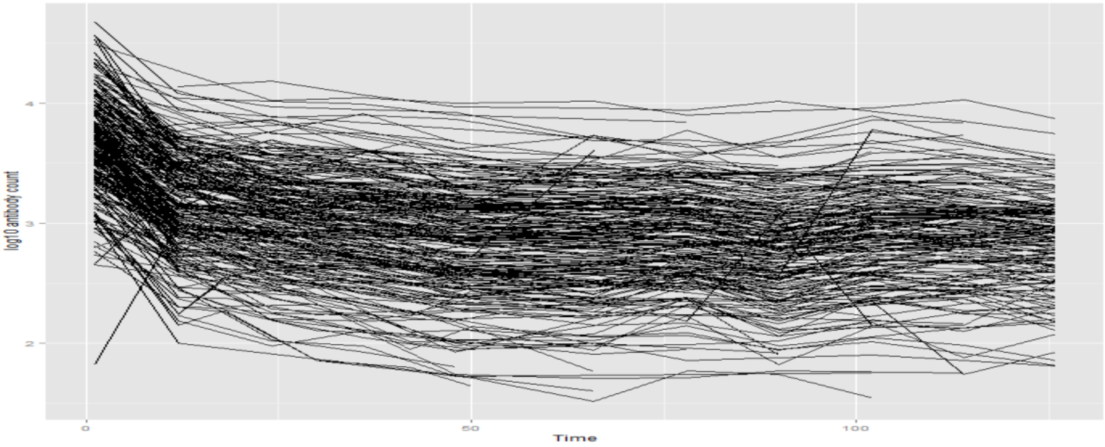

Figure 2: Longitudinal profiles of the log$_{10}$Antibody count.

There is missingness observed in our data set. We try to explore the missingness observed based on a logistic regression, by modeling the probability of observing a drop out based on the last observed antibody count. We do not observe a significant effect of the observed drop out on the probability of drop out. So it is quite possible that the missingness observed is random. . In practice it is never possible to confirm if the missingness is exactly at random. Since we look at the likelihood based models, where we use the observed data to make our parameter estimates, we could proceed with the assumption of missing at random as opposed to missing completely at random.

Statistical Analysis

We first look at the system of differential equations for the age independent model. We obtain three models, as have been described before. The similar kind of analysis has been done using MONOLIX, in 2012 by Andraud, et. al, where they had described the three time scales of antibody dynamics. Here we look at using the R package, *saemix* to estimate the population parameters from the data that we have at hand and compare the estimated parameter (only one realization) using the starting values as from the previous paper through the Non Linear Mixed Effect Models. We assume a random effect for all the parameters, and that the random effects are independent of each other, which is a very strong assumption to make.

All the estimates are obtained, using 5 chains and 5000 MCMC iterations.

Complete Model

Parameter	μ_a	μ_s	μ_l	A_0	φ_l	φ_s
Estimate	0.65	0.05	9.30E-07	3.80	1.80	0.38

We also obtain an AIC value of -1579.85 by the method of importance sampling. We see that the estimates of the short lived plasma cell life time are (μ_s^{-1}) which is approximately 20 months, and that of the long lived plasma cells is really huge. This is similar to the findings of the previous paper. The life time of the plasma cells are (μ_a^{-1}) about 1.5 months. Thus we see that the life spans of the short lived plasma cells are much longer than the antibodies. Also the rate of production of long lived plasma cells is much higher than the short lived plasma cells. The observed AIC is larger than what was observed in the past paper. So there exists scope for improvement of the estimates.

Asymptotic Model

Parameter	μ_a	μ_s	A_0	φ_l	φ_s
Estimate	0.66	0.05	3.79	1.84	0.37

We obtain an AIC value of -1580.06 by the method of importance sampling. We see that the estimates of the short lived plasma cell life time are (μ_s^{-1}) which is approximately 20 months. This is similar to the findings of the previous paper as well as the Complete Model. The life time of the plasma cells are (μ_a^{-1}) about 1.5 months. Thus we see that the life spans of the short lived plasma cells are much longer than the antibodies when the life span of long lived plasma cells is considered to be infinite. Also rate of production of long lived plasma cells is much higher than the short lived plasma cells. The observed AIC is larger than what was observed in the past paper. So there exists scope for improvement of the estimates.

Plasma Cell Driven Kinetic (PCDK) Model

Parameter	B_s	B_l	μ_s	μ_l
Estimate	2.8	0.83	0.08	8.10E-06

We obtain an AIC value of -1509.64 by the method of importance sampling. Here the estimates of the short lived plasma cell life time are (μ_s^{-1}) which is approximately 12 months. The estimated parameter values are quite close to the estimates in the original paper. We however obtain an AIC value which is lower than that was observed originally. Unlike the previous models that were discussed here, in the PCDK model, the production rate of long lived plasma cells is lower than that of short lived plasma cells, which is in line with the original paper.

In general, all the three models show a good consistency between the individual predictions and observations. The asymptotic model, like in the original paper has the lowest AIC value, but the PCDK model has a better fit in this case. However in general the final estimates depend heavily on the starting values of the parameters. The only advantage of the SAEM algorithm is the use of random sampling to estimate the maximum likelihood estimates. Thus the chances of being stuck in local minima are less.

Parameters	*Complete Model*	*Asymptotic Model*	*PCKD Model*
$\boldsymbol{B_s}$			2.800
$\boldsymbol{B_l}$			0.830
$\boldsymbol{\mu_s}$	0.050	0.050	0.077
$\boldsymbol{\mu_l}$	9.30E-07		7.10E-06
$\boldsymbol{\mu_a}$	0.650	0.660	
$\boldsymbol{\varphi_s}$	0.380	0.370	
$\boldsymbol{\varphi_l}$	1.800	1.840	
$\boldsymbol{A_0}$	3.800	3.790	
AIC- Importance Sampling	-1579.850	-1580.060	-1509.638

Table 2: Comparison of the different Age independent models

The diagnostic plots for the asymptotic model are presented below.

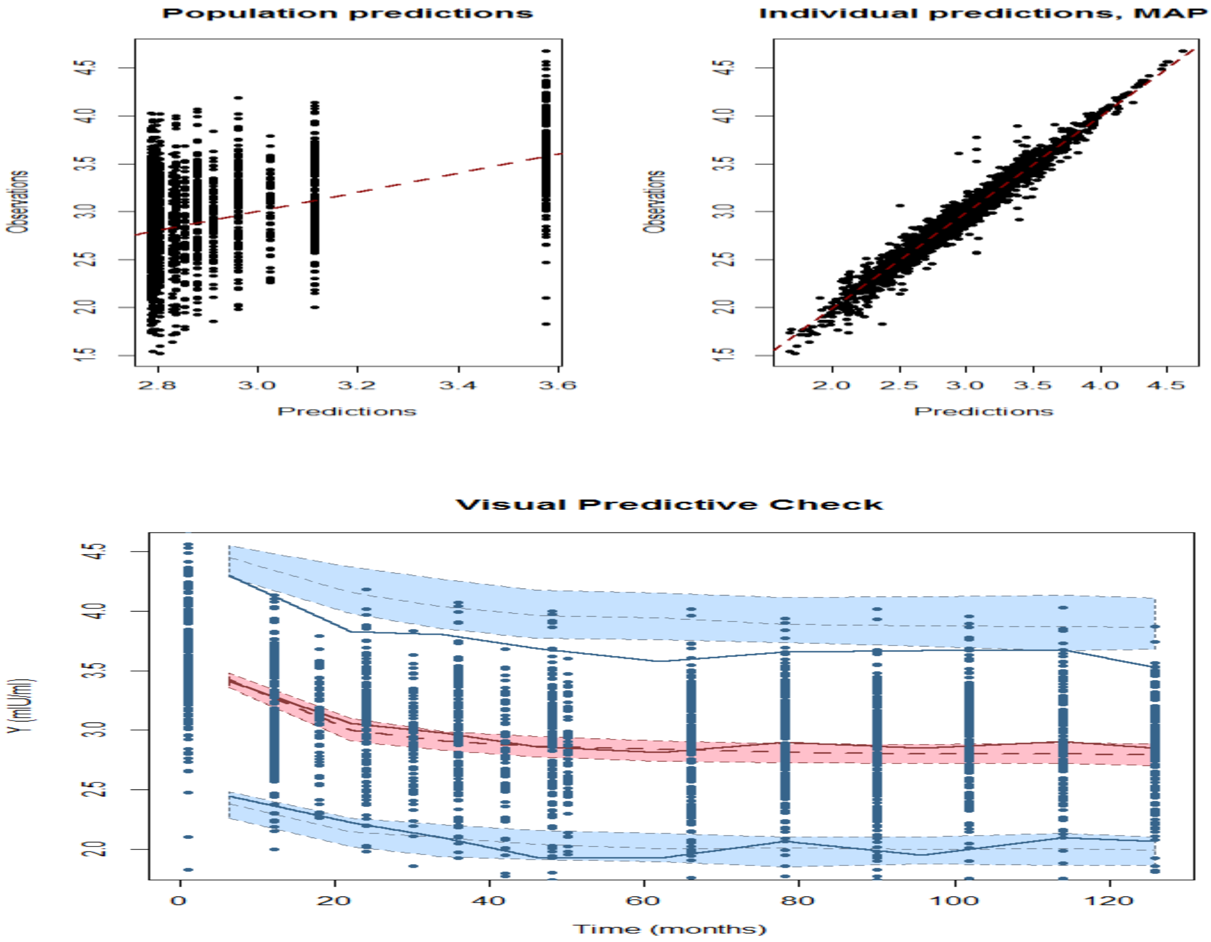

Figure 2: Predictions computed with the population parameters versus the observations (top left), and plot of the predictions computed with the individual parameters versus the observations (top right), prediction intervals around the boundaries of the selected interval (bottom).

The visual predictive checks include the prediction intervals around the boundaries of the selected interval as well as around the median (50th percentile of the simulated data). The top most intervals do not indicate a very good fit for the data unlike the other intervals.

The individual plots for 12 subjects have been shown in the following diagram.

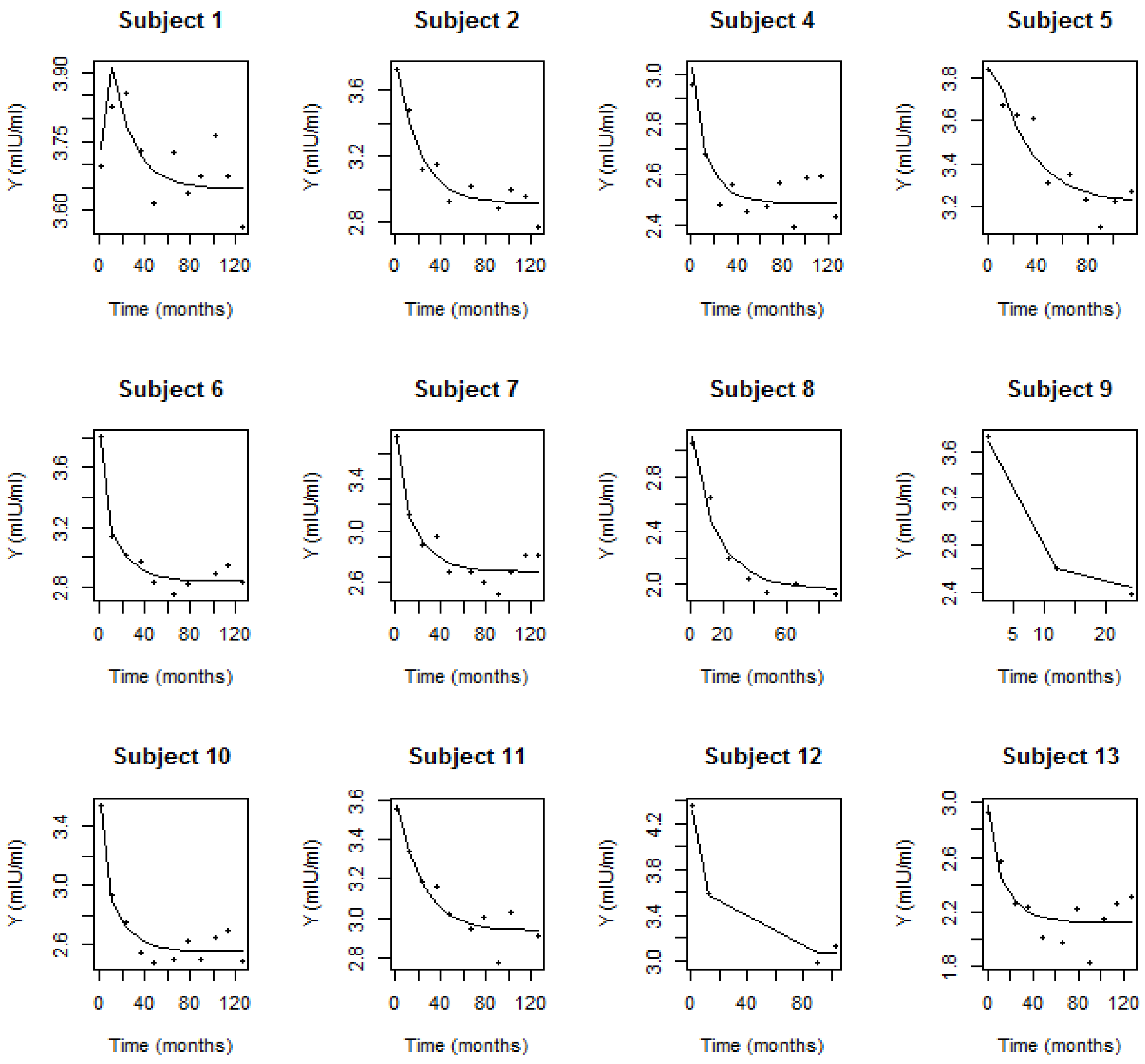

Figure 3: Individual plot for subject 1 to 12, using individual level parameters.

We now look at the models for the age dependent rates and summarize our findings below. We did not have any starting values in mind, so we started randomly and we proceed till we can minimize the AIC values.

Model 1: Exponential Decay and Constant Production Rate

Parameters	b_l	$\mu_l(0)$	η_l	b_s	$\mu_s(0)$	η_s
Estimates	1.000	0.12	0.027	4.100	1.80E-06	0.015

The calculated AIC value for the model is -1434.629 after a series of changing starting values, and it did not increase further. The terms , b_s and b_l indicate the plasma cell production rate and it is as per the ideas from the original paper that short term plasma cell production rate is much higher than the long term. Also the parameters η_l *and* η_s are positive, indicating that the age dependent decay process exists, increasing with increase in age. This may happen just by chance so formal statistical tests need to be performed to come to such a conclusion. These values have been obtained as a part of the optimization process of the likelihood function.

Model 2: Polynomical Decay function of Time and Constant Production Rate

Parameters	b_l	$\mu_l(0)$	η_l	β_l	b_s	$\mu_s(0)$	η_s	β_s
Estimates	1.000	0.055	0.004	0.024	0.003	1.7E-05	0.002	5.3E-06

The calculated AIC for this model is -1317.638, which is much less than the exponential decay rate model above. The age dependent kinetic parameters, η_l and η_s are smaller than in case of the previous model. The positive value indicates a progression in the decay rate with increase in age. Similarly like the previous model, the estimates were obtained as a part of the optimization process, and formal statistical tests need to be done to check its significance.

Model 3: Generalization of Model 1, with an Exponential Function defining the Production Rate

Parameters	b_l	$\mu_l(0)$	η_l	b_s	$\mu_s(0)$	η_s	γ_l	γ_s
Estimates	1.100	0.043	0.110	0.230	0.060	0.0005	0.008	1.7E-07

We calculate the AIC value for this model as -1552.51. The age dependent kinetic parameters, η_l is larger than the previous two models that have been studied but the η_s is smaller than in case of the previous model. The positive value indicates a progression in the decay rate with increase in age. We also see that the age dependent kinetic parameter for the production rate of the short lived plasma cells is very small and almost close to zero. However for the long lived plasma cells it is significantly larger, thus suggesting that long lived plasma cells are produced more with the increase in age, than the short lived plasma cells. Similarly like the previous models, the estimates were obtained as a part of the optimization process, and formal statistical tests need to be done to check its significance.

Model4: Generalization of Model 2, with an Exponential Function defining the Production Rate

Parameters	b_l	$\mu_l(0)$	η_l	β_l	b_s	$\mu_s(0)$	η_s	β_s	γ_l	γ_s	α_l	α_s
Estimates	1.30	0.06	0.003	8.1E-04	0.23	1.0E-05	0.001	3.4E-06	0.039	0.036	0.76	0.22

We calculate the AIC for this model as -1591.24. The age dependent kinetic parameters η_l and η_s for the decay are much smaller than the case where we assume the constant rate of production. Thus we see that rate of decay increases with an increase in age of the person. Similarly for the age dependent production rates, we see that the rate of production increases with age. The values for the long lived and the short lived plasma cells are almost similar. Similar to the above model, we see that long lived plasma cell production is much more than short lived plasma cells.

Parameters	Model1	Model2	Model3	Model4
b_l	1.0000	1.0000	1.1000	1.3000
$\mu_l(0)$	0.1200	0.0550	0.0430	0.0600
η_l	0.0270	0.0040	0.1100	0.0030
β_l		0.0240		0.0008
b_s	4.1000	0.0030	0.2300	0.2300
$\mu_s(0)$	1.8000E-06	1.7000E-05	0.0600	1.0000E-05
η_s	0.0150	0.0020	0.0005	0.0010
β_s		5.3000E-06		3.4000E-06
γ_l			0.0080	0.0390
γ_s			1.7000E-07	0.0360
α_l				0.7600
α_s				0.2200
AIC- Importance Sampling	-1434.6290	-1317.6380	-1552.5100	-1591.2400

Table 3:Comparison of the 4 scenario's of the PCDK model for the age dependent case.

The AIC value for the Model 4, considered is the smallest, and we there use this model for making the diagnostic checks.

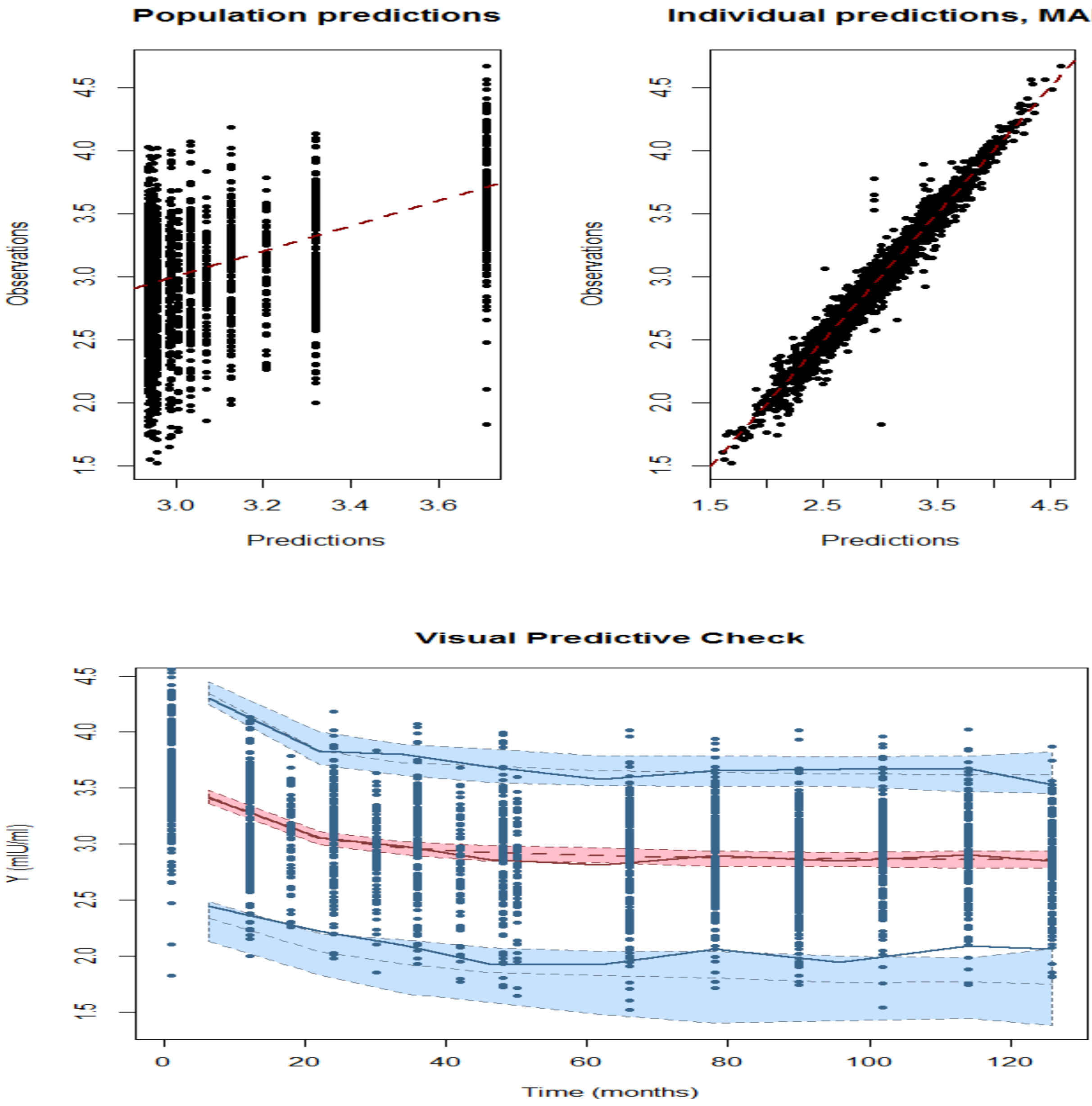

Figure 4: Predictions computed with the population parameters versus the observations (top left), and plot of the predictions computed with the individual parameters versus the observations (top right), prediction intervals around the boundaries of the selected interval (bottom).

The visual predictive checks include the prediction intervals around the boundaries of the selected interval as well as around the median (50th percentile of the simulated data). The bottom most intervals do not indicate a very good fit for the data unlike the other intervals. The visual predictive checks look much better than the age independent models.

The individual plots for 12 subjects have been shown in the following diagram.

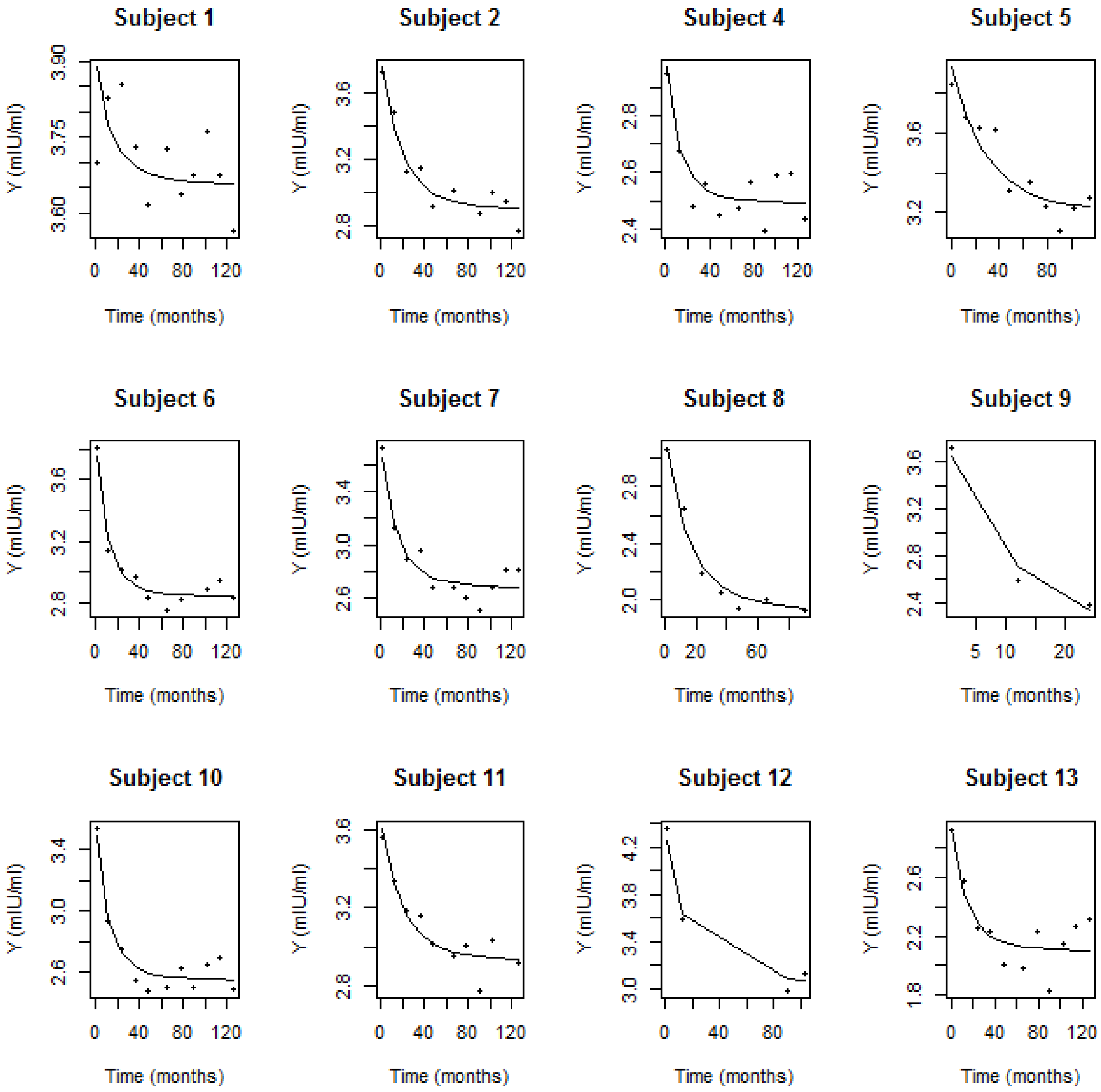

Figure 5: Individual plot for subject 1 to 12, using individual level parameters.

Conclusion and Discussion

In this report, we discussed the various aspects of Functional Data Analysis as a broad field of study. One of the major advantages of functional data analysis is that, it can be used to study many realistic physical and biological systems. Systems are generally described best by a set of differential equations, which may be ordinary, partial, stochastic, or algebraic. These systems finally result in a mostly non linear function in practice, some having closed form analytical expressions, but that's not always a necessity. The various methods that exist for the above broad field have been described. Non Linear Mixed Effect models are one of the most popular analysis tools that exist for such studies. The concepts of Inverse Models for systems are also coming up in statistical literature.

In particular we have looked at the system of Immunological activity in out body, post Hepatitis A vaccination. The underlying assumptions, on which the mathematical models were developed, come from the paper by Amana, et. Al, 2010, where it has been described that , initially there are antibody produced, followed by short lived plasma cells, and finally the long lived plasma cells, with an increasing life span for each of them respectively. We also considered two kinds of models, one assuming that the decay rates for the antibodies and plasma cells in our body are constant. However we know that our body immunity decreases with age, so we also look at an age dependent decay rate for antibodies and plasma cells and the models have been developed by Prof. Olivier Lejeune.

In the age independent case, we obtain analytical closed form expressions, for the antibody presence in our blood. In case of the age dependent models, analytical closed form expressions exist only for the Plasma Cell Driven Kinetic Model, where it is assumed that the antibody life span is much shorter than that of the plasma cells. The number of parameters in the complete model is far too many, and hence was not described in this report. In the asymptotic model, the expression for the antibody count could be expressed as an integral equation.

The Non Linear Mixed Effect models were fitted to the data with analytical closed form expressions, using the *saemix* package in R, which uses a Stochastic Version of the EM algorithm to come up with parameter estimates using the method of Importance Sampling. On a comparative basis, with NLMIXED package in SAS, the PCDK model for the age independent case took over 75 hours for convergence, whereas *saemix* could do the same job in 30 minutes.

The final model, that we selected based on the lowest AIC among the list of models that were considered assumed, a polynomial decay rate, thus the decay increases with time, since the coefficients are all positive. The production rate for the antibodies is considered to be exponential.

The power of the exponential is negative, thereby indicating that the production decreases with time.

In our case the *saemix* package could not estimate the Fischer Information matrix from the data, giving an error of “non conformable arguments” and as described by the builder of the package, Emanuelle Comets it could be possibly because of “a lot of random effects being present in the model”. Thus in our report, we present estimates of the parameters only, and we could not perform a statistical test for the significance of the estimated parameter from the data. However we performed visual predictive checks, which indicate a good fit for the data. As a general note, whenever we deal with problems of non linear optimization, the estimates largely depend on the starting value which is often a source of trouble. As for the integral equation, we did not use an approximation by the commonly used Taylor series expansion, since the presence of integral may inflate the error out of proportions. We did try to perform a non linear least square estimate calculation; however we experienced problems with starting values, and the Hessian matrix used in the approximation as becoming singular.

In regards to software packages that exist to solve these kinds of problems, a discussion has been presented. The Functional Mixed Effect model (*fme* package) approach using inverse problem methods to estimate parameters of an expression have certain drawbacks, like it cannot handle linear PDE’s or data of hierarchical data. Even *saemix,* cannot handle functions which do not have a closed form expression.

The question of dealing with the functional forms not existing, in certain examples have been looked at by a Bayesian approach by some researchers. However there is no existing methodology to handle situations when an expression/function is expressed only as an integral. Thereby no software packages exist, specifically for such kind of analysis. Due to the restricted time span of this thesis, I explored the methods and problems that exist in studying the antibody dynamics, and the mathematical challenges that are faced in this process. Thereby developing general methodology to solve all the existing problems remain an open question, for further research in terms of statistical methodology as well as in terms of computational software’s being available. Currently I am trying to use the procedure of bootstrapping to come up with bootstrap confidence interval for all the model parameters that were used. The results would be available in the sometime, due to the computational time involved in the computational process, and thereby we could guess something about the parameter significance level.

References

1: Ahmad,R. and Gray,D.(1996).Immunological memory and protective immunity: understanding their relation. Science: Apr 5; 272(5258):54-60.

2: Amanna,I.J. and Slifka,M.K. (2010).Mechanisms that determine plasma cell lifespan and the duration of humoral immunity. Immunol Rev 236: 125–138.

3: Andraud, M., Lejeune, O. et. al (2012). Living on Three Time Scales: The Dynamics of Plasma Cell and Antibody Populations Illustrated for Hepatitis A Virus. PLoS Computational Biology 8(3): e1002418.

4: Comets,E., Lavenu,A. and Lavielle,M. (2013). **saemix:** Stochastic Approximation Expectation Maximization (SAEM) algorithm. (http://cran.r-project.org/web/packages/saemix/index.html). Accessed on (15/08/2013).

5: Davidian, M. and D. Giltinan (1995). Nonlinear Models for Repeated Measurement Data. New York: Chapman and Hall.

6: Demicheli,V. and Tiberti,D.(2003). The effectiveness and safety of hepatitis A vaccine: a systematic review. Vaccine Jun 2;21(19-20) :2242-5.

7: Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society Series B 39, 1–38.

8: Ghosh,S.K. (2010). Non Linear Mixed Effect Models involving ODE's. 2010-11 Program on Semiparametric Bayesian Inference Applications in Pharmacokinetics and Pharmacodynamics, SAMSI, Research Triangle Park, NC, USA. (http://legacy.samsi.info/201011/pkpd/GhoshNLME-ODE-talk4in1.pdf). Accessed on (08/08/2013).

9: Goyal,L, and Ghosh,S.K. (2006). Statistical Inference For Non-linear Mixed Effects Models Involving Ordinary Differential Equations. Thesis Report. North Carolina State University,USA. (http://repository.lib.ncsu.edu/ir/bitstream/1840.16/4757/1/etd.pdf). Accessed on (10/08/2013).

10: Hagenbuch,N. (2011). A Comparison of Four Methods to Analyse a Non-Linear Mixed-Effects Model Using Simulated Pharmacokinetic Data–Thesis Report. ETH Zurich. (http://stat.ethz.ch/research/mas_theses/2011/hagenbuch.pdf). Accessed on (10/08/2013).

11: Hoppenstead,F.(1997). Mathematical Theories of Populations: Deomgraphics, Genetics, and Epidemics. Philadelphia: Society for Industrial and Applied Mathematics.

12: Irving,G.J., Holden,J., Yang,R., and Pope,D.(2012). Hepatitis A immunization in persons not previously exposed to hepatitis A. The Cochrane Database Systematic Reviews July; 7:CD009051.

13: Kuhn, E. and Lavielle, M. (2005), Maximum likelihood estimation in nonlinear mixed effects models. Computational Statistics and Data Analysis 49, 1020–1038.

14: Lindstorm,M.J. and Bates,D.M. (1990).Nonlinear Mixed Effects Models for Repeated Measures Data. Biometrics Vol. 46, No. 3 (September), pp. 673-687.

15: Levine, R. and Casella, G. (2001).Implementations of the Monte Carlo EM algorithm. Journal of Computational and Graphical Statistics 10, 422–439.

16: Molenberghs, G. and Verbeke, G. (2005).Models for discrete longitudinal data. New York: Springer.

17: Mieleitner, J. and Reichert, P. (2006).Analysis of the transferability of a biogeochemical lake model to lakes of different trophic state. Ecological Modelling 194, 49-61.

18: Park, S., et. al (2007). Statistical analysis of the dynamics of antibody loss to a disease-causing agent: plague in natural populations of great gerbils as an example. Journal of the Royal Society, Interface Feb 22;4(12):57-64.

19: Pillai, G. C., Mentre,F. , and Steimer,J.L. (2005).Non-linear mixed effects modeling - from methodology and software development to driving implementation in drug development science. Journal of Pharmacokinetics and Pharmacodynamics *32*, 161C183.

20: Pinherio,J.C. and Bates,D.M. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. Journal of Computational and Graphical Statistics Vol. 4, No. 1 (March), pp. 12-35.

21: Ramsay, J. O. and B. W. Silverman (2002). Functional Data Analysis (First ed.). New York: Springer.

22: Ramsay, J. O. and B. W. Silverman (2005). Functional Data Analysis (Second ed.). New York: Springer.

23: Rice, J.A. (2004). Functional and Longitudinal Data Analysis: Perspectives on smoothing. Statistica Sinica 14,613-629.

24: Soetaert,K., Petzoldt,T. and Setzer,R.W. (2010). R Package **deSolve**: Solving Initial Value Differential Equations.

25: Soetaert,K. and Petzoldt,T. (2010). Inverse modelling, sensitivity and Monte Carlo analysis in R using package **FME**. Journal of Statistical Software, 33(3):1–28.

26: Soetaert,K. and Petzoldt,T. (2012). **FME:** A Flexible Modelling Environment for Inverse Modelling, Sensitivity, Identifiability, Monte Carlo Analysis.(http://cran.r-project.org/web/packages/FME/index.html). Accessed on (15/08/2013).

27: Tornoe, C. W. et al. (2004).Non-linear mixed-effects pharmacokinetic/pharmacodynamic modeling in NLME using differential equations. Computer Methods and Programs in Biomedicine,76(1), 31-40.

28: Ullah, S. and Finch, C.L. (2013). Applications of functional data analysis: A systematic review.BMC Medical Research Methodology, 13:43.

29: Vajda,S. ,Rabitz, H., Walter,E., and Lecourtier,Y. (1989). Qualitative and Quantitative Identifiability analysis of non linear chemical kinetic models. Chemical Engineering Communications 83(1):191-219.

30: Verbeke, G. and Molenberghs, G. (2010). Advanced Modeling Techniques. Universiteit Hasselt Course Notes, Hasselt.

31: Wang,L. (2007). Estimating Nonlinear Mixed Effect Models by the Generalized Profiling Method and its Application to Pharmacokinetics- Thesis Report, McGill University, Canada.(http://digitool.library.mcgill.ca/webclient/StreamGate?folder_id=0&dvs=1378835313394~150). Accessed on (08/08/2013)

32: Wang, J. (2007). EM algorithms for nonlinear mixed effects models. Computational Statistics and Data Analysis 51, 3244 – 3256.

33: Wei, G.C.G. and Tanner, M.A. (1990). Calculating the content and the boundary of the highest posterior density region via data augmentation. Biometrika 77, 649–652.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Testing non-autonomous models of antibody dynamics by parametric fitting of data on HAV vaccination: Exploratory study

Richting: **Master of Statistics-Biostatistics**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Bakuli, Abhishek

Datum: **11/09/2013**