University of New Mexico
UNM Digital Repository

Mathematics & Statistics ETDs | Electronic Theses and Dissertations

6-23-2015

Nonparametric Bayes approach for a semi-mechanistic pharmacokinetic and pharmacodynamic model

Yan Dong

Follow this and additional works at: https://digitalrepository.unm.edu/math_etds

Recommended Citation
Dong, Yan. "Nonparametric Bayes approach for a semi-mechanistic pharmacokinetic and pharmacodynamic model." (2015).
https://digitalrepository.unm.edu/math_etds/67

This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at UNM Digital Repository. It has been accepted for inclusion in Mathematics & Statistics ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.

Yan Dong

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Ronald Christensen , Chairperson

Michele Guindani

Gabriel Huerta

Erik Erhardt

Nonparametric Bayes approach for a Semi-Mechanistic Pharmacokinetic and Pharmacodynamic Model

by

Yan Dong

MS in Statistics, University of New Mexico, 2011

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2015

©2015, Yan Dong

Dedication

To my family and friends.

Acknowledgments

I would like to express sincere gratitude to my advisor, Dr. Michele Guindani for his constant guidance and support throughout my entire study. I also thank to Dr. Ronald Christensen, Dr. Gabriel Huerta and Dr. Erik Eahardt for their insight and comments on my dissertation.

It has been my pleasure to study in the department and at UNM. I really appreciate all the people who have made the resource available to me, without which I could not complete the study. Finally, I would like to thank Yong and my family for their endless support and encouragement for me.

Nonparametric Bayes approach for a Semi-Mechanistic Pharmacokinetic and Pharmacodynamic Model

by

Yan Dong

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2015

Nonparametric Bayes approach for a Semi-Mechanistic Pharmacokinetic and Pharmacodynamic Model

by

Yan Dong

MS in Statistics, University of New Mexico, 2011

Ph.D., Statistics, University of New Mexico, 2015

Abstract

Both frequentist and Bayesian approaches have been used to characterize population pharmacokinetics and pharmacodynamics(PK/PD) models. These methods focus on estimating the population parameters and assessing the association between the characteristics of PK/PD and the subject covariates. In this work, we propose a Dirichlet process mixture model to classify the patients based on their individualized pharmacokinetic and pharmacodynamic profiles. Then we can predict the new patients' dose-response curves given their concentration-time profiles. Additionally, we implement a modern Markov Chain Monte Carlo algorithm for sampling inference of parameters. The detailed sampling procedures as well as the results are discussed in a simulation data and a real data example. We also evaluate an approximate solution of a system of nonlinear differential equations from Euler's method and compare the results with a general numerical solver, ode from R package, deSolve.

KEY WORD: Dirichlet Process; Nonparametric Bayes; Pharmacokinetics; Pharmacodynamics; Ordinary differential equations.

Contents

Contents

List of Figures

List of Figures

Chapter 1

Introduction

The development of drugs is time-consuming and costly. DiMasi $(2003, 2001)$ reports that it takes about $10 - 15$ years to develop one new medicine from the time it is discovered to when it is available for treating patients. The average cost to research and develop each successful drug is estimated to be $\$80$ million to $\$1$ billion. This amount of money includes the cost of thousands of failed compounds. For every $5000 - 10,000$ compounds that enter the research and development pipeline, only one receives approval. For many years, the U.S. Food and Drug Administration (FDA) has encouraged the development of computational modeling and simulation to improve the efficiency in developing safe and effective drugs. Therefore the technological advances in different fields related to drug development is largely demanding to optimise current drug development practices.

The drug development consists of a preclinical and clinical component. The preclinical process includes all early research on a large number of different compounds tested in animals to gather safety and efficacy information. The process takes approximately $3 - 6$ years. By the end, the researchers hope to have a promising candidate drug test in people. The clinical study component is divided into three phases. In phase I, the drug is tested in healthy humans ($20 - 80$ participants) for physiological compatibility. The goal in this

phase is to determine the drug's most frequent side effects and, often, how the drug is metabolized and excreted. In phase II, the drug's therapeutic effects are investigated and the goal is to obtain preliminary data on whether the drug works in human who present a certain disease or condition. Phase II trials sometimes involve hundreds of patients. In phase III, the drug is tested in a larger sample of the population. The researchers gather more information about safety and drug efficacy. They try to further understand the drug behavior from different groups of patients and dosing regimens.

Pharmacokinetic and pharmacodynamic properties of a drug are of utmost importance in clinical trials. Simply stated, pharmacokinetics (PK) describes the time course of a drug in the body, whereas pharmacodynamics (PD) describes the study of the pharmacological effects of a drug. If we are able to better understand the relationship between PK and PD, we can expect to improve our knowledge of the underlying background mechanisms, and eventually optimize the dosing regimen of a treatment to develop a personalized drug therapy.

In this work, we propose to use the clustering property of Bayesian nonparametric infinite mixture models to borrow strength in the estimation of a semi-mechanistic PK/PD model. First, we establish a coherent probabilistic model which connects the individualized PK and PD model. Second, we classify the patients into several homogeneous groups on the basis of PK and PD profiles. More importantly, we can predict a new patient's PD profile based on its PK profile. The classification, estimation and prediction are achieved in a single model framework. We also review two traditional approaches for PK and PD model, nonlinear mixed effect model and the simple Bayesian hierarchical model. The estimation of the nonlinear mixed effect model depends on the likelihood function. The second one has been used to incorporate prior information in the model fitting. We discuss a modern Markov Chain Monte Carlo sampling algorithm for posterior inference in our model. In the end, we evaluate our model in a simulation data and a published clinical

trial data. In the simulation data, we also use Euler's method to linearize the system of nonlinear ordinary differential equations (ODEs) and compare its results with a general numerical ODE solver, ode, from an R package (deSolve).

The structure of the thesis is as follows. In chapter 2 we introduce PK and PD as well as the models describing PK/PD process. Chapter 3 presents two existing statistical approaches for fitting PK/PD models: nonlinear mixed effect models and Bayesian models. In chapter 4 we present the nonparametric Bayesian framework and discuss a modern Markov chain Monte Carlo algorithm that we will use later. In chapter 5 we discuss our proposed approach and provide a detailed MCMC sampling algorithm. The relevant results from the simulation data and a real data are also presented. In chapter 6 we present the conclusions and discuss future research work.

Chapter 2

Population pharmacokinetic and Pharmacodynamic (PK/PD) Models

The whole PK/PD modeling process includes careful data management (import, process, visualization of the time-course data), development of a PK/PD model, estimation of parameters using statistical methods, prediction and extrapolation beyond the existing data. Population pharmacokinetics studies the variability of plasma drug concentrations in a certain population after a standard dosage administered (Aarons 1991). The aim of PK is quantifying the variability within the population and accounting for it in terms of patient covariates, such as age, sex, disease state. Population pharmacokinetics can identify the measurable factors that cause changes in the dose-concentration relationship and the extent of these changes. It helps us better understand how these factors affect the absorption, distribution and elimination of the drug.

Population pharmacokinetics and pharmacodynamics describes the relationship between drug response and concentration of a drug and the variability of inter- and intra-subjects in a population. The population PK/PD models consist of two parts. The first

part describes a population PK model and we can obtain the PK parameter estimates for each individual. The second part is the corresponding PD model. There are primarily two ways to perform the population PK/PD modeling: simultaneous and sequential. In the simultaneous modeling, both PK an PD estimates are obtained at the same time. In the sequential approach, PK parameters are estimated for each individual first and then PD parameters are estimated based on PK information as a known covariate in the PD modeling. In our approach, we use the simultaneous approach to estimate the PK and PD parameters. In this chapter, we introduce the data-based non-compartmental model, compartmental model and physiology-based model for PK, steady-state and non steady-state PD models.

2.1 Pharmacokinetic models

Pharmacokinetics studies how the body affects a specific drug over time. It attempts to discover the fate of a drug from the moment that it is administered till it is completely eliminated from the body. The whole process includes the mechanism of absorption, distribution as well as the chemical changes of the drug, i.e. metabolic enzymes. Absorption is the process that involves drug movement from site of entry into bloodstream. After adminstration, a drug will be distributed itself into all of body's compartments and tissues that it is able to. The time it takes for this to occur is called distribution phase. Volume of distribution can be estimated based on sampling blood concentration after dosing with the assumption that the drug uniformly distributed throughout the body. Metabolism is the process by which a drug is chemically inactivated, i.e. broken down by enzymes, so that it can be excreted from the body. The excretion is the process that a drug is removed from the site of action and eliminated from the body. After a dose of drug is administered, the body begins to eliminate the drug by hepatic metabolism, renal excretion or both. The aim

of PK modeling is to explain and characterize the variability of these processes based on the observations over time. Next we review three different pharmacokinetic models, data-based non-compartmental model, compartmental model and the physiology-based model.

2.1.1 Data-based non-compartmental models

The data-based non-compartmental model is the simplest approach to measure the drug exposure and explain the variability across doses and subjects. The parameters include maximum plasma concentration, C_{max}, time to reach C_{max}, T_{max}, area under the plasma concentration-time curve AUC which determines the drug exposure over a period of time, total clearance CL which describes how quickly drugs are eliminated, metabolized or distributed throughout the body, total volume of distribution V_c etc. The non-compartmental model is highly dependent on the estimation of the AUC, calculated by numerical integration e.g. a trapezoidal rule. If we have a smooth line for concentration versus time or an equation for C_t from a pharmacokinetic model we could slice the area into vertical segments. The total area AUC is the sum of all the segments. In calculus, the total area is given by 2.1,

$$AUC = \int_0^\infty C_t dt \tag{2.1}$$

The thinner the segments are, the closer the trapezoids reflect the actual shape of the concentration-time curve. A sum of exponential terms are often used to describe the drug concentration-time profiles as follows,

$$C_t = A \cdot \sum \exp^{-\lambda_i t} \tag{2.2}$$

where C_t, A, λ_i are drug concentration over time, coefficient constant and terminal rate respectively (more explanations in 2.5 and 2.7).

In summary, the non-compartmental models are widely useful for data description and interpolation, Gabrielsson (2012). They can also be used to show that two formulations are

the same in the development of new drugs by calculating AUC and C_{max}. For example, the manufacturer may study the drug initially with capsules but may wish to market a tablet to patients as the latter trend to have greater stability. To do so, the manufacturer has to show that the two formulations are bio-equivalent which means that the two formulations have the same concentration-time profile within an acceptable tolerance region. In this case, calculate the C_{max} and AUC under both formulations for n patients. Bio-equivalence is declared if the confidence interval for the ratio of means of test/reference for both AUC and C_{max} are entirely within the interval $(80\%, 125\%)$, US Food and Drug Administration, Bioavailability (2000). In addition, they are often implemented in physiologically-based pharmacokinetics and pharmacodynamic models Gillespie (1991). However, they are difficult to do extrapolation, Aarons (2005). Because their parameters do not have the physiological interpretations. Thus it is difficult to predict how the concentration-time profiles change when the underlying physiology changes. This problem can partly be addressed by adopting the compartmental models.

2.1.2 Classical compartmental models

In order to obtain insight from a certain study, (e.g. biological systems), scientists divide the objects of scientific interest into smaller conceptual units until the underlying mechanisms become apparent (a basic logic in science). This basic principle also works for the pharmacological phenomena. A class of small conceptual units have been developed, known as compartments. The organism where the drug goes is thought of as a system of interconnected pools, compartments. In order to understand the whole mechanism, the researchers need to study the single compartment and the passage of the drug between them. For simplicity, we will describe the one- and two-compartmental models and discuss advantages and limitations of using compartmental modeling approach.

One-compartmental model is assumed to comprise only a single compartment repres-

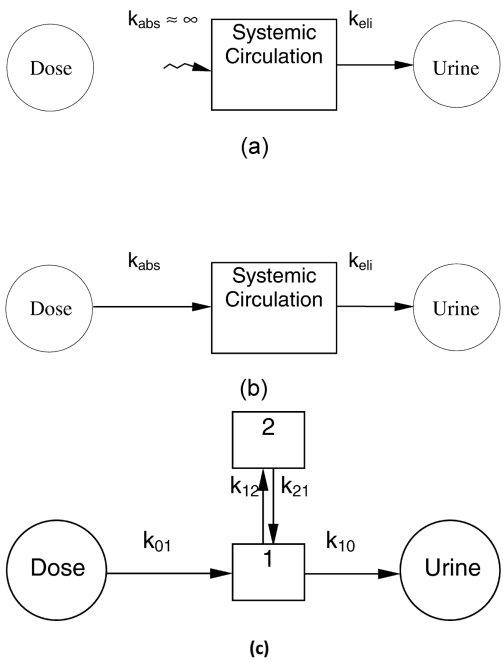

Figure 2.1: Examples of one and two compartment models modified from Holz et al. (2001) (a) Bolus/pill administration (b) First order absorption (c) Two compartment model

enting the systemic circulation. Figure 2.1.2 shows an one-compartmental model for a single input pathway and single elimination pathway, a with bolus/pill administration and b IV administration with first-order absorption. The uptake source and the place of elimination process are usually displayed as circles while the compartments of the organism are symbolized by squares or rectangles. where k_{abs} and k_{eli} are constants indicating the rate of absorption and elimination respectively. The main interest is the drug concentration as a function of the time after administration. The "into" and "out" flow of the process in the compartment can be described mathematically. The "into" process can be described by

$$\frac{dA(t)}{dt} = -k_{abs} \cdot A(t), \tag{2.3}$$

and the "out" process can be expressed by

$$\frac{dC(t)}{dt} = k_{abs} \cdot F \cdot A(t) - k_{eli} \cdot C(t), \tag{2.4}$$

where $A(t)$ is the amount of drug still remaining to be absorbed and F is the absolute bio-availability (describing how much drug reaches the circulation system after administration, i.e. the bio-availability of an intravenous drug dose is assumed to be 100%). The mathematical solution of the differential equations 2.3 and 2.4, with the initial value $A(0) = Dose$ and $C(0) = 0$ is

$$C(t) = \frac{F \cdot Dose \cdot k_{abs}}{V_d \cdot (k_{abs} - k_{eli})}\left(e^{-k_{eli}t} - e^{-k_{abs}t}\right), \tag{2.5}$$

where V_D is the volume of circulation compartment. Given the values of Dose, V_D, k_{abs}, k_{eli} and F, the drug concentration in the blood over time can be easily obtained.

The two-compartmental models resolve the body into a central compartment and a peripheral compartment (see Figure 2.1.2 (C)). They are assumed that the central compartment includes the tissues that are highly perfused such as blood, heart, lungs, liver, kidneys and brain. The peripheral compartment comprises poor-perfused tissues such as muscle, fat and skin. After drug administrated into the central compartment, the drug is assumed to distribute between that compartment and the peripheral compartment. However, the drug does not achieve instantaneous distribution between the two compartments. Here we present a simple two-compartmental model which is described as follows. The system of ordinary differential equations is:

$$\begin{aligned} \frac{dC_1(t)}{dt} &= -(k_{10} + k_{12})C_1(t) + k_{21}C_2(t), \\ \frac{dC_2(t)}{dt} &= k_{12}C_1(t) - k_{21}C_2(t), \end{aligned} \tag{2.6}$$

where C_1 and C_2 represent the amount of drug in the central and peripheral compartment,

respectively. The solution for bolus administration with $C_1(0) = Dose$ and $C_2(0) = 0$ is

$$\begin{aligned} C(t) &= A \cdot e^{-\alpha \cdot t} + B \cdot e^{-\beta \cdot t} \\ &= \frac{F \cdot Dose}{V_D(\alpha - \beta)} \left[(\alpha - k_{21}) e^{-\alpha t} + (k_{21} - \beta) e^{-\beta t}) \right] \end{aligned} \tag{2.7}$$

with

$$\alpha \cdot \beta = k_{12} \cdot k_{10} \tag{2.8}$$

and

$$\alpha + \beta = k_{12} + k_{21} + k_{10} \tag{2.9}$$

where k_{10}, k_{12}, k_{21} are the elimination rates constant from compartment one, the inter-compartmental flow rate respectively. With a simple algebra, these rate constants can be solved easily.

The model becomes more complicate when we start considering multiple compartments in the system, such as catenary and mammillary compartmental structure see Figure 3 in Holz (2003). A mammillary model consists of a central compartment interacting with a number of peripheral compartments surrounding it. The catenary model comprises of a chain of interconnected compartments. The time course of the concentrations will always follow a sum of exponentials under a certain assumption. However, it may need more complex techniques to solve the equations.

Nowadays, the compartmental models are still widely used in various area. Kreuer (2014) proposed a three-compartmental pharmacokinetic model extended with an additional lung compartmental and clearance to measure drug concentration in patient's breath. Yamazaki (2015) proposed a one-compartmental PK model to describe oral and subcutaneous profiles of anaplastic lymphoma kinase inhibitors.

2.1.3 Physiologically-based Pharmacokinetics models (PBPK)

In the 1930s, Teorell ($1937a, b$) provided a set of equations for uptake, distribution, and elimination of drugs from the body. These papers are regarded as providing the first physiological model for drug distribution. However, computational methods were not available to solve the sets of equations at that time. The focus shifted to the simple models such as data-based non-compartmental and one- or two- compartmental models. Because they have fewer number parameters and these parameters do not correspond directly with a specific physiological compartment. For the next thirty years, PK modeling focused on these simpler descriptions with exact solutions. However, with the availability of computers and numerical integration algorithms, it regains the interest in physiology-based model from 1970s. By 2010, hundreds of publications used PBPK models and some companies' business based on their expertise in PBPK. There is also a growing interest in applying PBPK models for the discovery and development of drugs, Lupfert 2005.

PBPK models are compartmental models like the classical pharmacokinetics compartmental model, but the compartment here represents the actual tissue and organs. In general, the concept of PBPK is to use mathematical equations to describe relevant physiological, biochemical process which determines the pharmacokinetic behavior of a compound. PBPK model is structured to compose the relevant physiological compartments. Each compartment often represents a single organ or tissue. These compartments are interconnected via the blood circulation loop. The mass-balance equations for each compartment describe the rate of substance change within it.

Figure 2.1.3 illustrates the generic stucture of building a PBPK model. It includes the monitoring of the drug concentration in core tissue, fluid and organs (arterial and venous blood), liver (the main metabolising organ), kidney (for renally excreted drugs). Within each compartment, mathematical equations are derived from the law of mass transfer. Linear or nonlinear differential equations are the most common description of the pharmacokinetic processes. The PBPK models have two groups of parameters: drug-specific and physiological parameters. Typical drug-specific parameters include metabolism rate and plasma protein binding constants. The physiological parameters include regional and tissue blood flow rates, volumes of blood and different tissues parameters etc. An example of a system of ordinary different equations are described for a real world problem Craigmill (2003). This model is to see how a chemical injected into a body of an animal spread into all the organs and how the concentration of chemical changes over time in each organs. The diagram for the model is shown 2.1.3, with a complicate differential equations followed 2.1.3. See more details and explantation in the original paper.

The PBPK modeling has attracted considerable attention in pharmacological and toxicological research (Grass and Sinko, 2002). There are also more applications in the dose estimation (Johnson 2005) and drug-drug interactions (Chien, Monhutsky et al. 2003). Another important area of PBPK application is the drug discovery and development in Lupfert and Reichel (2005), Latz (2009). Jones and Rowland-Yeo (2013) presents a comprehensive tutorial for PBPK. Huang and Rowland (2012) discussed the role of PBPK modeling in regulatory review. We refer the interested reader for PBPK models to Reddy (2005).

In summary, the use of PBPK modeling to maximize the clinical potential of drugs has been accepted in pharmacokinetics. But it requires intense resources to generate the data on the various parameters in the model. The mathematical complexity of the model and computationally intensive limit the application of PBPK modeling. Therefore, user-friendly software would enhance the widespread use as well as the knowledge of

physiology and biochemical process, especially in different disease states.

2.2 Pharmacodynamic models

Pharmacodynamics (PD) refers to the relationship between drug concentration at the site of action and the resulting effect, including the intensity of therapeutic and adverse effects. The magnitude of a drug response at the action site is determined by the amount of drug binding to a certain type of receptor. More drug is at the action site, the stronger intensity of a drug's effect is. PD model quantifies the relationship between dose and response. There are two basic assumptions in PD models. The first one is that the measured plasma concentration is proportionally related to the concentration at the effect site. Ideally, the concentrations should be measured at the effect site, the action site or biophase where the interaction with the corresponding receptor system occurs, but this is not possible for most drugs. However, the concentrations could be easily accessed in body fluids like plasma or blood. The plasma concentrations are frequently used to establish the relationship between a dose of drug and response.

The second assumption is about the drug effect. The drug effect can be defined as any drug-induced change in a physiological parameter when compared to the baseline value. The baseline is the values of the same physiological parameters in the absence of drug dosing. Baseline values do not necessarily have to be constant but can change, i.e. as a function of day. Moreover, "effect" has to be clearly separated from "efficacy". Efficacy is the sum of all therapeutically beneficial drug effects and is the most relevant target parameter in clinical trial. In practice, efficacy is difficult to quantify and thus instead use the easily accessible surrogate markers as effect parameters. But it needs to present evidence that the effect parameters used correlates with the desired efficacy. In practice, the measured response could be both continuous and categorical, i.e. blood pressure,

cure/not cure,none, mild moderate and severe. In those cases logistic regression models and survival analysis are applied to describe the probability of the events. The relationship between concentration and the logit of the probability of the event is typically modeled as a linear or E_{max} function. For a dichotomous longitudinal variable the probability of an outcome $P(t)$ may be estimated based on

$$L(t) = E_0 + Slope \times C(t) \qquad P(t) = \frac{\exp^{L(t)}}{1 + \exp^{L(t)}} \tag{2.10}$$

where $L(t)$ and E_0 are the total and underlying effects on the logistic scale. In this section we introduce the PD models in steady-state and non steady-state conditions.

2.2.1 Pharmacodynamic models for steady-state situations

When the concentration of the agent at the action site are constant and the PD parameters are time-invariant, the system is said to be kinetically at steady state. The steady-state condition can be reached with long-term IV infusions or multiple-dose regimens. Several basic PK/PD models have been used to describe concentration-effect relationship, such as fixed effect model, linear model, log-linear model, sigmoid E_{max} model Meibohm (1997). The frequently used one is the sigmoidal E_{max} model, Mager (2003),

$$E(t) = \frac{E_{max} \times C(t)^r}{EC_{50}^r + C(t)^r} + E_0 \tag{2.11}$$

where E_{max} is the maximum effect that can be achieved by the drug in the investigated system and EC_{50} is the drug concentration that results in half of the maximum effect. EC_{50} is inversely related to the potency. γ is the sigmoidicity factor that determine the steepness of the relationship but is in many cases not statistically significant from 1. Among effect and log-concentration relationship,

- If $r = 1$, it is for a hyperbolic curve.

- If $r > 1$, the relationship becomes steeper and will eventually approach a step-function.
- If $r < 1$, it is for a smoother curve.

One reason for the popularity of the E_{max} model is that the function asymptotes to an upper limit of stimulation or inhibition by a drug on a system. However, often there are situations when sufficienty high concentrations can not be achieved to estimate E_{max} and simplification can be made where fewer parameters are estimated. When the concentrations are much smaller than E_{50}, the E_{max} model collapses to a linear model ($\gamma = 1$) or a power function with coefficient Slope as shown in 2.12,

$$E(t) = E_0 + Slope \times C(t)^{\gamma} \tag{2.12}$$

Another issue is that the underlying E_0 is not always constant over the drug period. For example, the effect variable may vary because of an underlying disease, such as fluctuations in glucose in the event of diabetes in blood pressure. The model complexity can increase with increased availability of data and knowledge of the underlying system. For example, there may be feedback mechanisms that regulate the measured variable, such as the influence of insulin on glucose levels. In addition, after drug administration, the drug effect delay are frequently observed. There are multiple issue to affect the delay, such as slow distribution to the effect site, active metabolite formation, signal transduction and other mechanisms.

2.2.2 Pharmacodynamic models for non-steady-state situations

Under non-steady-state conditions, the time course of plasma concentration and effect dissociate. Thus, to fully characterize the time course of drug action, PK and PD have to be adequately linked to predict the relationship of PD effect versus plasma concentration. Next we review two basic attributes in the integrated PK/PD models.

Direct Link vs. Indirect Link

If drug concentration in plasma occurs rapidly enough at the effect site, which means the temporal delay negligible, the concentrations are directly proportional in between over time. The effect and plasma concentration lack any detectable hysteresis and may be directly linked. However, in many cases, the concentration at the site of action may lag behind that in plasma; then no direct link can be established. This is usually manifested by a hysteresis between plasma concentration and effect. See an example in Suri et al. (1997). The extent of hysteresis is dependent on the degree of delay between the concentrations in plasma and at the effect site. This can be resolved by introducing a hypothetical effect compartment representing active drug concentration at the effect site. Linking the effect compartment to the kinetic model with negligible mass of drug into the effect compartment.

Time Variant vs. Time Invariant

In PD model, most parameters are assumed to time invariant, i.e. E_{max} and EC_{50} stay constant over time. However, when time-variant PD occurs, a specific model for the involved process of tolerance is required. The tolerance is defined as a decrease in drug effect over time, despite constant drug concentrations at the effect site, and is characterized by a clockwise hysteresis loop in a plot of effect versus drug concentration. In an E_{max} model, tolerance is usually modeled as a time-dependent decrease of EC_{50} if receptor desensitization is assumed.

In summary, under non-steady and steady-state, there are several pharmacodynamic models have been used to describe the relationship between drug concentration and effect. The preferred models depend on many factors, including $1)$ the type of the drug used $2)$ the response to be measured, $3)$ the effect seen after administration of drug, $4)$ the linearity in

the effect-concentration curve, $5)$ the characteristics of parameters whether represent the underlying process and mechanism.

2.3 Population PK/PD Models

The population approach is often employed to indicate a paradigm that attempts to define important PK and PD differences and extract this information from complex data. The ultimate goal of PK/PD analysis is to establish guidelines for individualizing dosage regimen. It consists of quantifying the mean and variance of PK/PD population parameters as well as the intra- and inter- subject variability. In addition, it is also possible to investigate the factors or covariates, such as age, weight, gender, which may help to distinguish the difference among individuals or subgroups of the population.

2.3.1 Mechanism-Based PK/PD models

The PK data analysis is often considered routine and straightforward, but major physiological insights have derived from basic physiology principles. The time course of drug concentrations in a relevant biological fluid, i.e. unbound plasma concentration, C_p, are typically represented by a mathematical function:

$$C_p(t) = f(\theta_{PK}, Dose, t) \tag{2.13}$$

where θ_{PK} is a vector of PK parameters determined by model fitting, $Dose$ represents the amount of medicine, t is the time course.

If plasma concentrations are assumed to be proportional to biophase concentrations, then these expressions are served as known information in PD models.

$$E(t) = f(\theta_{PD}, C_p, t, X) \tag{2.14}$$

where $E(t)$ is the pharmacological response over time and X represents a vector of drug-independent system parameters, i.e. age. Both equations 2.13 and 2.14 may be explicit for some simple systems. In general, the population PK/PD approach imposes the distributional assumption on the individual-specific PK and PD parameter vectors, θ_{PK} and θ_{PD}.

$$\theta_{i,PK} \sim N(\mu_{pk}, \Sigma_{pk}) \qquad \theta_{i,PD} \sim N(\mu_{pd}, \Sigma_{pd}) \tag{2.15}$$

where $N(\cdot, \cdot)$ denotes a p-dimensional multivariate normal distribution with mean and covariance matrix. The main interest is to use the concentrations as a primary predictor to construct a realistic model for drug effect that allow the efficacy and various covariates to be explored. Nielsen and Friberg (2013) promoted more extensive use of modeling and simulation to describe time courses of antibiotic drug effects in animals and patients. Their review summarize the value of PK/PD modeling and provided an overview of the characteristics of available PK/PD models of antibiotics. Dong (2014) developed a population PK/PD model for mycophenolic acid in paediatric renal translate recipients. They used a two compartmental model with a transit compartment for PK modeling. For the PD model, a non-linear relationship between dose and acid exposure was described by a power function.

The objective of a PK/PD modeling is not just to describe the data sets of the sample of individuals but also used to simulate which concentration and effect and evaluate the variability of drug response for the future patients when different doses are given. These simulation and prediction can lead to optimised dosing recommendation. It is often said that "all models are wrong but some are useful". In order to define whether a population PK/PD model is useful and valid in clinical study. A number of evaluation and validation models have been performed such as goodness-of-fit models, bootstrap analysis, visual predictive evaluation. Cock (2010) reviewed a few validation models of PK/PD modeling in paediatric clinical research. Burns (2014) proposed a population PK/PD model of

caffeine by using visual analogue scale and evaluated both simultaneous and sequential PK/PD modeling. Model validation were performed using diagnostic plots and visual predictive check plots.

The "mean" functions in population PK/PD models are often expressed by a system of differential equations describing the kinetic and dynamic process in the body. Most often, the analytic solutions of these equations do not exist. However, there are several numerical solvers available, such as Runge-Kutta methods, adams methods and backward differentiation formula. The choice of these approaches is on the base of the characteristics of the systems, i.e. stiffness. An ODE system is called stiff if the state variables change on a wide variety of time scales, including changing very rapidly as well as changing very slowly. Solving "stif" ODEs may request the special methods. An excessive amount of computing time is required because it takes time to use very small time steps to maintain stability. In addition, solving a system of ODEs depends on the "good" choice of the starting values of parameters. The common packages of ODE solvers are available in the programming languages i.e. FORTRAN (ODEPACK solver), C (CVODE solver), MATLAB (ode45/ode15s solver) and R (deSolve package). In our work, we use R package, deSolve for solving PK and PD ODEs and also implement a self-written R code of Euler's method for PD ODEs.

2.3.2 Overview of softwares of population PK/PD modeling

There are various softwares available for PK, PBPK and PBPKPD modeling in drug development processes. They are Phoenix WinNonlin, P-PHARM, PHEDSIM, MEDICI-PK, Modkine, PDx-MC-PEM and JGuiB. These are well user friendly software for not only simple PK/PD model but also for population PK/PD models. WinNonlin is popular because it includes extensive libraries for PK and/or PD and PK/PD models and it also provides tools for table generation, scripting and data management. Roccheti (2009)

provided a PK/PD analysis for estimating PK/PD parameters and computing the expected tumor growth curves was carried out by WinNonlin $V.3.1$.

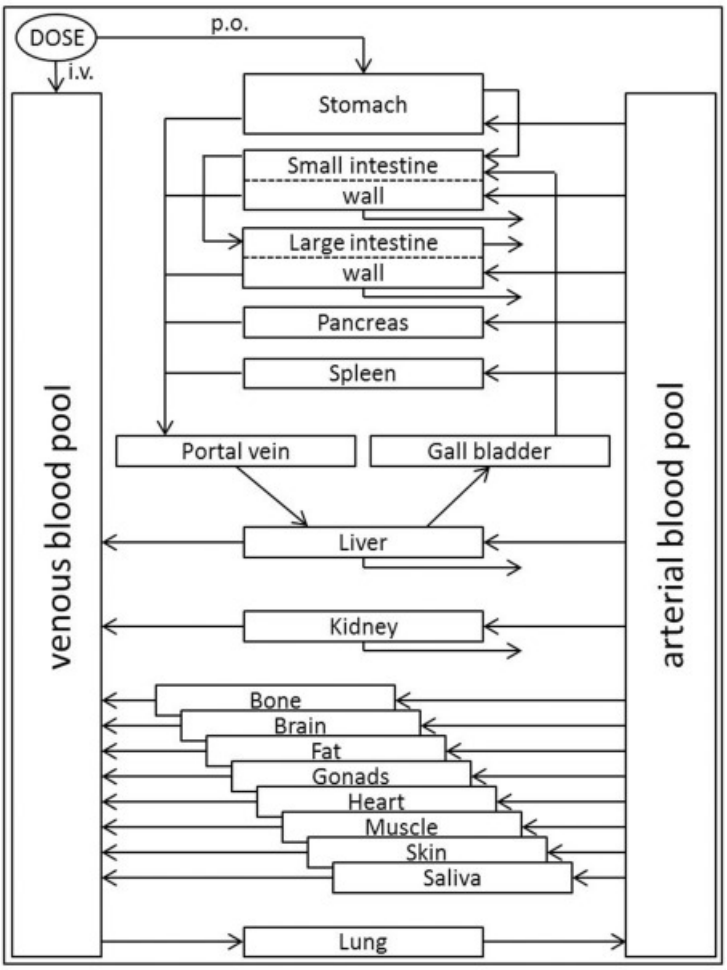

Figure 2.2: A generic structure of the whole-body PBPK-model Dickschen (2012). The model organism is built by compartments, each typically representing a single organ defined by its physiological volume. Organs are interconnected via respective blood flows which occur, except for pulmonary circulation, from the arterial blood pool to the venous blood pool thus accounting for inter-compartmental mass-transfer. Application of substances can be defined as intravenous (i.v.), per oral (p.o.), or into any desired compartment. In addition to clearance events in intestinal wall, liver, and kidney, metabolism processes can be implemented into any compartment. Transport processes that significantly influence a compounds PK may be inserted between compartments.

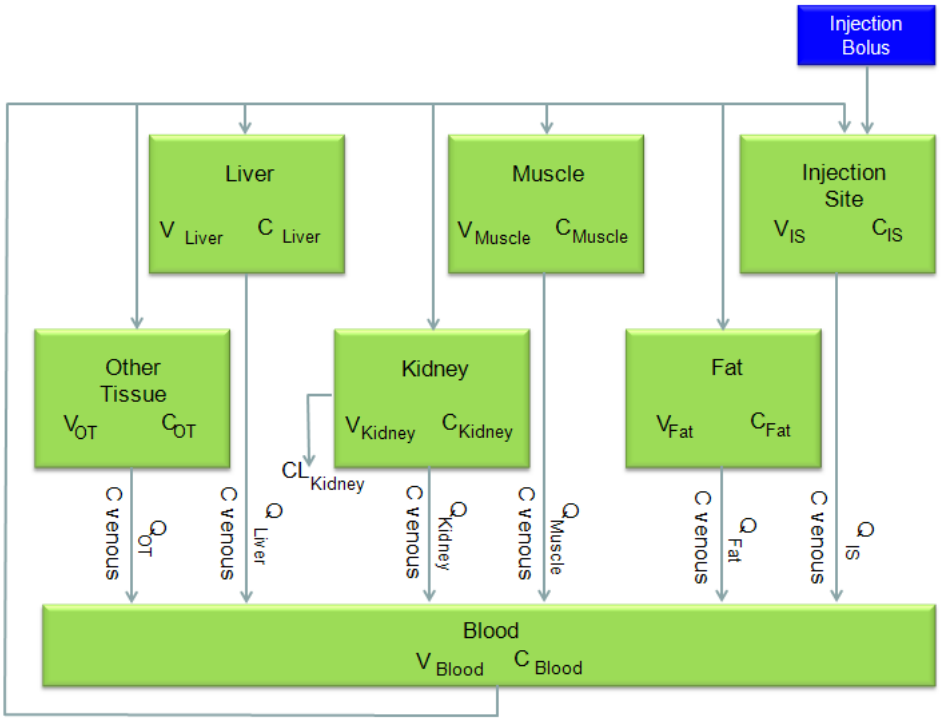

Figure 2.3: A diagram of PBPK model structure Craigmill (2003).

$$\frac{dC_{muscle}}{dt} = \frac{[C_{blood} - (C_{muscle} / R_{muscle})] \cdot Q_{muscle}}{V_{muscle}}$$

$$\frac{dC_{IS}}{dt} = \frac{[C_{blood} - (C_{IS} / R_{IS})] \cdot Q_{IS} + [k_{slow}(1-F)BW \cdot DOSE) + (k_{fast} F \cdot BW \cdot DOSE)]}{0.5}$$

$$\frac{dC_{liver}}{dt} = \frac{[C_{blood} - (C_{liver} / R_{liver})] \cdot Q_{liver}}{V_{liver}}$$

$$\frac{dC_{fat}}{dt} = \frac{[C_{blood} - (C_{fat} / R_{fat})] \cdot Q_{fat}}{V_{fat}}$$

$$\frac{dC_{OT}}{dt} = \frac{[C_{blood} - (C_{OT} / R_{OT})] \cdot Q_{OT}}{V_{OT}}$$

$$\frac{dC_{kidney}}{dt} = \frac{[(C_{blood} - (C_{kidney} / R_{kidney})) \cdot Q_{kidney} - C_{kidney} CL_{kidney}]}{V_{kidney}}$$

$$\frac{dC_{blood}}{dt} = \frac{[(C_{liver} / R_{liver}) Q_{liver}] + [(C_{kidney} / R_{kidney}) Q_{kidney}] + [(C_{muscle} / R_{muscle}) Q_{muscle}] + [(C_{fat} / R_{fat}) Q_{fat}] + [(C_{OT} / R_{OT}) Q_{OT}]}{V_{blood}}$$

$$- \frac{C_{blood} Q_{blood}}{V_{blood}}$$

Figure 2.4: An example of a system of differential equation for each compartment Craigmill (2003).

Chapter 3

Literature review on statistical approaches for population PK/PD models

In the field of biomedical applications, data usually consists of repeated measurements on individuals under varying experimental conditions. For example, in pharmacokinetic studies several blood samples are taken on participants over a period of time following the administration of a drug. Those participants can be considered as a random sample from a population of interest. The measured response is often not nonlinear with the varying experimental condition with the parameters. The model is fitted to data sets from different individuals. The main interest is to obtain inference on both population and individual characteristics, and their variability.

Two common approaches are considered to fit the population PK/PD data. The first approach is to fit all data sets simultaneously in one simple model (simultaneous method). In other words, PK and PD parameters are estimated at the same time. The second approach

is to fit a model for the PK data first, then model the PD outcome conditional on the PK data and/or estimates ("sequential" method). Zhang et al. (2003) uses the simulation data to compare the performance of a simultaneous method with the sequential method. They validated the results with respect to computation time, estimation precision and inference. In the end, they concluded that the computation time of the sequential method is less and the estimates are more easily obtained. Thus in our work, we fit both PK and PD data simultaneously in our MCMC algorithm.

PK/PD have a precise administration in clinical trial study. The patients often receive different doses of drug. Blood samples are collected over a period of time. In our real data, paclitaxel was administered as a 3-hour infusion, with initial dose of $175mg/m^2$ every $3rd$ week. Dose adjustments were guided by hematological and nonhematological toxicity which resulted in a final dose range of 110 to 232 mg/m^2. Plasma concentrations were monitored on course 1 and course 3, with an average of 3.5 samples per patient and course.

There are several statistical methodologies to model the population PK/PD data. The simplest approach is called "naive-pooled dat". The "naive-pooled data" approach is to fit all individuals' data together without considering inter- and intra- subject variability. The "two-stage" approach fits each individual's data separately and then combines the individual parameter estimates. In the first stage, each individual PK/PD parameters are estimated from the individual PK and PD observations. In the second stage, the relationship between covariates and the parameters are explored. The population mean and variance of each parameter is derived. The third approach is the nonlinear mixed models. Nonlinear mixed models is a mixture of fixed and random effects. Fixed effect is often repeatable and the experimenter can directly manipulates. Random effect is the source of random variation. The fixed effects estimate the population coefficients but the random effects account for the individual difference in response. The fourth approach is Bayesian. which

incorporates the initial belief for the parameters in the model. Historically, the difficulty of implementing the Bayesian analyses in complex statistical models was the intractability of the numerous integrations. However, vigorous development of Markov chain Monte Carlo (MCMC) techniques facilitated such integration in the early 1990s. Additionally new advances in computing power have made such Bayesian analysis feasible. The nonlinear mixed model served as one of the first examples of this capability (Rosner and Muller 1994). We will provide a brief review of the important features of Bayesian inference. See Davidian and Giltinan (1995, Ch.8) for an introduction and Carlin and Louis (2000) for comprehensive coverage of modern Bayesian analysis. In this chapter, we review last two approaches for the PK/PD models. In section 3.1, we review nonlinear mixed-effect models with available programs. In section 3.2, we discuss the Bayesian models and available software.

3.1 Statistical Inference for Nonlinear Mixed-effect Models

Within the framework of nonlinear models, the main interest is focused on representing the mean function, or mean trajectory, describing the dynamic relationship between the response and explanatory variables, i.e. time. Most often, the means functions are described by a system of ordinary differential equations (ODEs). (We skip ODEs for now and discuss it more at the end of this chapter.)

The nonlinear mixed-effect model is the traditional approach for longitudinal data. The concept of nonlinear mixed-effect model (NLMEM) first appeared in (Sheiner, Rosenberg, Melmon 1972), modeling of individual PK for computer aided drug dosage. In 1977, the first case study used NLMEM shown up in Sheiner, Rosenberg Melmon. They estimated the population characteristics of PK parameters from routine clinical data. In late 1980s,

it was widely spread in statistical research. There are numerous new methodologies and computational techniques developed for these models in late 1990s, like NONMEN software (an IBM-specific software). The estimation methods for population parameters are the First Order method, Laplace approximation, and First Order linearization etc. Sheiner and Beal published three important papers and evaluated the methods for estimating population pharmacokinetic parameters. From 1985 to 1990, nonparametric, bayesian estimation of individual random effects given current estimates appeared, such as linearisation of the model around the current estimates of the random effect, Newton-Raphson iterative solution to a linear mixed effect estimation problem (Lindstrom and Bates (1990)).

Here we first introduce a simple version of nonlinear mixed-effect models and then review several inferential methods. Let y_{ij} denote the jth observed response, for ith subject, measured at time point t_{ij}, for $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n_i$. For example, in pharmacodynamic settings, y_{ij} is the absolute neutrophil count at time t_{ij} for subject i. The following methodology we present can be extended to a more general case with multiple covariates. A statistical model can be expressed as,

$$y_{ij} = f(\boldsymbol{\theta}_i, t_{ij}) + \epsilon_{ij}, \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, n_i \tag{3.1}$$

For an individual i, the intra-individual error ϵ_{ij} corresponds to the measurement uncertainty associated with the observed response at time point t_{ij}. The random errors are assumed to be independently distributed with zero mean and constant variance across all measurements,

$$E(\epsilon_{ij}) = 0 \quad \text{and} \quad Var(\epsilon_{ij}) = \sigma^2 \quad \text{for} \quad i = 1, 2, \cdots, m \quad \text{and} \quad j = 1, 2, \cdots, n_i \tag{3.2}$$

For simplicity, errors are assumed to be $i.i.d$, identically independent distributed. Although this assumption is very strong, it can be relaxed to the case that the variance of errors to be, $Var(\boldsymbol{\epsilon}_i) = \sigma_i^2 R_i$, where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})$, R_i is a $n_i \times n_i$ positive definite matrix. R_i may depend on other parameters. For simplicity, we consider R_i as an identity

matrix with the number of n_i elements. In equation 3.1, $f(\cdot)$ is the mean function describing the within-individual behavior. It depends on a vector of p parameters, $\boldsymbol{\theta}_i$, specific to individual i. $\boldsymbol{\theta}_i$s are assumed to from a common distribution with mean $\boldsymbol{\theta}$ and variance $\boldsymbol{\Sigma}_\theta$. Specifically we can write,

$$\boldsymbol{\theta}_i = d(\boldsymbol{\theta}, \boldsymbol{b}_i), \qquad \boldsymbol{b}_i \sim N(0, \boldsymbol{D}) \tag{3.3}$$

3.3 represents the individual behavior conditional on $\boldsymbol{\theta}_i$ and hence on $\boldsymbol{b}_i$, the random component in 3.3. In 3.3, we assume that the distribution of $\boldsymbol{b}_i|\boldsymbol{X}_i$ does not depend on all the covariates $\boldsymbol{X}_i$. All $\boldsymbol{b}_i$'s have a common multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{D}$. In our real data, the PK parameter vector $\boldsymbol{\theta}_i$ is $(CL_i, V_i^c, V_i^d, CL_i^d)$ representing the system clearance, volume of central compartment, volume of peripheral compartment, clearance of inter-compartment separately.

The joint density of the observed data $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m$ is,

$$\prod_{i=1}^{m} \int p(\boldsymbol{y}_i, \boldsymbol{b}_i|\boldsymbol{\theta}) d\boldsymbol{b}_i = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} p(\boldsymbol{y}_{ij}|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i) d\boldsymbol{b}_i \tag{3.4}$$

Ideally, the parameters involved in the NLME model can be estimated by maximizing the joint density of all parameters based on equation 3.4. If f is a linear function of parameter $\boldsymbol{\theta_i}$, the integral can be evaluated to obtain an analytic expression. Here we present a very simple example where the joint density of parameters can be formalized. We do not consider any covariates for now and set

$$\boldsymbol{\theta}_i = \boldsymbol{\theta} + \boldsymbol{b}_i, \quad E(\boldsymbol{b}_i) = \mathbf{0}, \quad Var(\boldsymbol{b}_i) = \boldsymbol{D} \tag{3.5}$$

The likelihood function assuming normality of the responses and random effects can be written as,

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{D}, \boldsymbol{\sigma}) &\propto \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} \frac{1}{D} e^{-\frac{1}{2\sigma^2}(y_{ij} - f(\boldsymbol{\theta}_i, t_{ij}))^2} |\boldsymbol{D}|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta})^T \boldsymbol{D}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta})} d\boldsymbol{\theta}_i \\ &= \prod_{i=1}^{m} (\frac{1}{\sigma^2})^{n_i/2} |\boldsymbol{D}|^{-1/2} \int e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij} - f(\boldsymbol{\theta}_i, t_{ij}))^2 - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta})^T \boldsymbol{D}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta})} d\boldsymbol{\theta}_i \end{aligned} \tag{3.6}$$

Even though the likelihood in 3.6 has a specific form. However, in most cases of NLME models, f is a nonlinear function of parameters $\boldsymbol{\theta}_i$. Nonlinearity means that the m integration in equations 3.4 and 3.6 can not be solved in a closed form. Thus, it is not possible to obtain the analytic solution. In practice, the classical approaches are based on approximation of the likelihood involving the linearization of the nonlinear model by using either Taylor's series expansion (Beal and Sheiner (1982)) or by applying Laplace's approximation to the likelihood (Wolfinger, 1993). Then estimate the parameters based on the approximated likelihood functions. Here we review several methods based on approximation of the likelihood according to the equations from 3.1 to 3.3.

3.1.1 Exact method

By exact methods, we mean methods which avoid approximations, such as exact maximum likelihood estimation (MLEs) or Bayes estimation. In order to obtain MLEs or bayesian estimate, one must evaluate a high dimensional integral which requires numerical methods, i.e. Monte Carlo. For MLE, the marginal density of y_i given the parameters b_i, θ is required,

$$p(y_i|\theta, b_i) = \int p(y_i|\theta, b_i)p(b_i)db_i \tag{3.7}$$

In most cases, f is nonlinear in the random effect b_i. So there is no closed form expression for this integral. Both Calculating MLEs or Bayesian estimates can be computationally difficult and time consuming especially when b_i is more than one-dimensional. Next we use the simplest model to introduce several approaches to approximate the function f.

3.1.2 First order methods

This method, developed by Beal and Sheiner (1982), is extensively used in pharmacokinetics. First order method is based on using a Taylor expansion to approximate the model function f, and maximizing the likelihood corresponding to the resulting approximation. The first-order Taylor expansion of $f_i(\cdot, b_i)$ around $b_i = 0$ gives the approximate model

$$y_i = f_i(\theta_i, t_{ij}) + \epsilon_i \approx f_i(\cdot, b_i = 0) + f_i'(\cdot, b_i = 0)b_i + \epsilon_i \tag{3.8}$$

where $f_i'(\cdot, 0)$ is the derivative of $f_i(\cdot, \cdot)$ with respect to b_i evaluated at $b_i = 0$. We can write down the approximate marginal distribution of y_i which is normal with expection $f_i(\cdot, 0)$ and variance $\sigma^2 I_{n_i} + f_i'(\cdot, b_i = 0)D[f_i'(\cdot, b_i = 0)]^T$. Maximizing the corresponding likelihood with respect to rest of parameters gives Beal and Sheiner estimates. This method gives desired results in some situations but gives significantly biased estimates, i.e. when $f_i(\phi, b_i)$ is significantly non-linear in b_i.

3.1.3 Conditional First-order method

The conditional first-order algorithm also uses a first order Taylor expansion, but it expands at both estimates of the random effects, $b_i = \hat{b_i}$ and and estimate of other parameters, denoted by a vector ϕ. Next we describe the algorithm.

Step 1, set the initial values D_0 and σ_0^2, the estimates of $\hat{\phi}$ and $\hat{b}_i$ are obtained by minimizing,

$$\sum_{i=1}^{m} \Big(\frac{1}{\sigma_0^2} ||y_i - f_i(\phi, b_i)||^2 + b_i' D_0^{-1} b_i \Big) \tag{3.9}$$

Step 2, expand the model $f_i(\phi, b_i)$ around $b_i = \hat{b}_i$ and $\phi = \hat{\phi}$,

$$y_i \approx f_i(\hat{\phi}, \hat{b}_i) + f_i'(\hat{\phi}, \hat{b}_i)(\phi - \hat{\phi}) + f_i'(\hat{\phi}, \hat{b}_i)(b_i - \hat{b}_i) + \epsilon_i \tag{3.10}$$

Then y_i approximately follows normal with different expectation $f_i(\hat{\phi}, \hat{b}_i) + f_i'(\hat{\phi}, \hat{b}_i)(\phi - \hat{\phi}) - f_i'(\hat{\phi}, \hat{b}_i)\hat{b}_i$ and same covariance as in first-order method. This becomes a linear mixed effect problem and $\hat{\phi}, \hat{\sigma}^2$ and $\hat{D}$ maximize the approximate likelihood.
Step 3, set $\sigma_0^2 = \hat{\sigma}^2$ and $D_0^2 = \hat{D}^2$, repeat step 1 and 2 until convergence.
See more details in Davidian and Giltinan's book and Wolfinger 1993 are good references on this approach.

3.1.4 Laplacian Method

The Laplacian method evaluates the exact marginal likelihood by using a second-order Taylor expansion of l_i around the empirical Bayes estimate $\mathbf{b}_i$. Let the first and second derivatives of l_i describe as

$$l_i' = \frac{\partial l_i}{\partial \mathbf{b}_i}$$
$$l_i'' = \frac{\partial^2 l_i}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}$$

Thus

$$\begin{aligned} l_i &= l_i(\hat{\mathbf{b}}_i) + l_i'(\hat{\mathbf{b}})(\mathbf{b}_i - \hat{\mathbf{b}_i}) + \frac{1}{2}(\mathbf{b}_i - \hat{\mathbf{b}_i})^T l_i''(\hat{b_i})(\mathbf{b}_i - \hat{\mathbf{b}_i}) \\ &= l_i(\hat{\mathbf{b}}_i) + \frac{1}{2}(\mathbf{b}_i - \hat{\mathbf{b}_i})^T l_i''(\hat{\mathbf{b}_i})(\mathbf{b}_i - \hat{\mathbf{b}_i}) \end{aligned} \tag{3.11}$$

Consequently, the integral in can be approximated by 3.11. See more details in Wolfinger (1993) and Davidian and Giltinan (1995).

3.1.5 Lindstrom and Bates algorithm

The Lindstrom and Bates algorithm can be derived using Laplacian approximation. The estimation algorithm use a combination of a penalized nonlinear least-square estimate

(PNLS) and a linear mixed-effect estimate (LME). They used a first-order Taylor expansion on the conditional estimates of the individual random effect. To simplify the notations, denote the covariance matrix for random effect $D^{-1} = \sigma^{-2}\Delta^T\Delta$. In the penalized nonlinear least-squares, the random effects $\mathbf{b}_i$ and the estimate of the fixed effect β based on the current estimate of Φ are obtained by minimizing the function as follows,

$$O_{PNLS} = \sum_{i=1}^{N}(\mathbf{y}_i - f_i(\mathbf{z}_i, \theta_\mathbf{i}))^T(\mathbf{y}_i - f_i(\mathbf{z}_i, \theta_\mathbf{i})) + \mathbf{b}_i^T\Delta^T\Delta\mathbf{b}_i \tag{3.12}$$

In order to update the estimate of Φ, the mean function $f(\cdot)$ is linearized using a first-order Taylor expansion around the current estimate of β and the estimate of $\mathbf{b}_i$. The approximate log-likelihood function for the estimation of Φ can be written as

$$logL(\theta, \sigma^2, \Delta) = -\frac{\sum_{i=1}^{N} n_i}{2} log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\Bigg\{log(I + \frac{\partial f_i}{\partial b_i^T}\Delta^{-1}\Delta^T\frac{\partial f_i^T}{\partial b_i^T})$$
$$+\Big[\mathbf{y}_i - f_i(\mathbf{b}_i, \theta_\mathbf{i}) + \frac{\partial f_i}{\partial \mathbf{b}_i}^T\hat{\mathbf{b}_i}\Big]^T\Big(I + \frac{\partial f_i}{\partial b_i^T}\Delta^{-1}\Delta^T\frac{\partial f_i^T}{\partial b_i^T}\Big)^{-1}\Big[\mathbf{y}_i - f_i(\mathbf{b}_i, \theta_\mathbf{i}) + \frac{\partial f_i}{\partial \mathbf{b}_i}^T\hat{\mathbf{b}_i}\Big]\Bigg\}$$

There are more methods besides our review i.e. adaptive Gaussian quadrature in SAS, stochastic approximation expectation maximization. Vonesh and Carter (1987) proposed the use of estimated generalized least squares and establish the asymptotic properties of the resulting estimates. An alternative method is the use of iteratively weighted generalized least squares. The MIXNLIN program also implements pseudo maximum likelihood and restricted maximum likelihood estimation by embedding the EM algorithm within a re-weighted generalized least squares routine. The expansion is either about 0 or about the empirical best linear unbiased predictor of the inter-individual random effects. All approaches started first from initial estimates set to the true values and second using altered values. Roe et al. (1997) and Duffull et al. (2005) provide a systematic comparisons of these population modeling and a summary of the estimation algorithms.

The success of the statistical techniques nowadays are directly related to the availability of reliable, efficient and user-friendly software for its application. In pharmacology, non-

linear mixed effects models are the most common method to describe pharmacokinetics and pharmacodynamic relationship. It is very useful for mapping out many different kinds of dose-response curves. In this section we briefly review 3 widely used packages.

The first one, NONMEN software was developed by Beal and Sheiner (1980) and has been widely used by practitioners to implement PK/PD data analysis. It performs maximum likelihood estimation based on several approximation of log-likelihood function, such as first-order, first-order conditional estimation methods and Laplacian method. This software has various attractive features thus it is generally regarded as the gold standard software for PK/PD modeling and fitting non-linear mixed effects models. NONMEN can fit the standard PK/PD compartment models. Moreover, it can handle the multiple dosing regimen in the model. More importantly, it can fit models expressed by the system of ordinary differential equations. In general, NONMEN is quite accurate, stable, flexible to fit PK/PD models.

The second one, R/S-plus package nlme can also be used to fit nonlinear mixed effect model. This package can not only fit two levels of random effect of PK/PD models but the multiple-level model non-PK/PD models. nlme also has many functions to be used for modeling checking, plotting as well as diagnostics. nlme itself cannot handle compartment models expressed by ODE's without closed-form solution. It requires using an ODE solver to these PK/PD models. **nlmeODE** package combines **nlme** with the **odesolve** package in R. The **odesolve** package provides an interface to the Fortan ODE solver Isoda, which can be used to solve initial value problems for systems of first-order ODE's. The computation times are usually significantly longer using **nlme** with **nlmeODE** compared to NONMEN.

The NLMIXED procedure in SAS is also widely used to fit non-linear mixed effects models. PROC NLMIXED uses integration approximation methods to optimize the objective function. Thus it can be viewed as giving the "exact answer" to the optimization problem. The advantage of NLMIXED procedure is that it allows users to specify the like-

lihood function. In SAS, it is very easy to perform exploratory analysis, manipulate data and perform diagnostics. The output can be directed to files in a variety of formats using the SAS Output Delivery System. The NLMIXED procedure is quite sensitive to starting values and parameterizations of the model.

3.2 Bayesian Hierarchical Models

A Bayesian hierarchical model can be considered as the analogous of the mixed effect models in the frequentist statistical approach. Historically, the implementation of Bayesian analyses in complex statistical models is not computationally feasible. However, the development of MCMC techniques in the early 1990s have made posterior inference feasible. MCMC techniques can produce samples from the relevant posterior distributions, from which any desired function of the parameters of interest may then be approximated. For population PK/PD models, MCMC algorithms may be more difficult to implement because of the complexity of the mean functions and the nature of the data. Thus, the implementation in available software such as WinBUGs and PKBugs may be more challenging. Rosner and Muller (1994) and Wakefield et al. (1994) provided the first examples of a Bayesian approach in the analysis of PK/PD data. In this section, we present a complete Bayesian approach for a population PK/PD model originally developed by Lunn et al. (2002) and briefly mention the available softwares.

3.2.1 Bayesian E_{max} based model for population PK/PD model inference

First we present a simple Bayesian model for inference on population PK-PD model. Let y_{ij} denote the PK response for ith subject at observed time point j, $i = 1, 2, \cdots, m$;

$j = 1, \cdots, n_i$. Let θ_i denote the p-dimensional vector of PK parameters for individual i and σ^2_{pk} denote the PK variance for the measurement noise. The sampling distribution is assumed to be normal,

$$p(y_{ij}|\boldsymbol{\theta}_i, \sigma^2) \sim N(f(\boldsymbol{\theta}_i, t_{ij}, D_i), \sigma^2_{pk}), \tag{3.13}$$

where the y_{ij}s are either concentration or log-concentrations measurement depending whether normality or log-normality is the most appropriate assumption for the data. $f(\cdot)$ is a function of individual-specific parameters $\boldsymbol{\theta}_i$, observed time t_{ij} and the dosing history, D_i. Assign a p-dimensional multivariate normal prior distribution. The individual-specific PK parameter vector $\boldsymbol{\theta}_i$ is then,

$$p(\boldsymbol{\theta}_i) \sim N_p(Z_i\boldsymbol{\mu}, \Sigma) \tag{3.14}$$

where Z_i is a $p \times q$ covariate-effect design matrix for individual i, μ is a population-level vector of coefficients and Σ ($p \times p$) represents the population-level variance-covariance matrix.

The model is completed by assuming a prior on the population-level parameters, σ^2_{pk}, μ and Σ,

$$\sigma^2_{pk} \sim IG(a, b); \qquad \mu \sim N_q(\eta, H); \qquad \Sigma \sim IW(R, \rho) \tag{3.15}$$

where $IG(\cdot, \cdot)$ and $IW(\cdot, \cdot)$ denote the inverse-gamma and inverse-Wishart distributions, respectively. The hyper-parameters a, b, η, H, R and ρ are fixed based on available prior information.

If PK modeling is of interest, then relevant inference focuses on the population level parameters, σ^2_{pk}, μ and Σ. However, where the drug effect has been measured, the PK model becomes of intermediate interest, to predict drug concentrations in a certain "effect compartment". These concentration measurements are then used as a primary predictor

(input) to construct a realistic model for drug effects that allow the relationship between efficacy and various covariates to be explored.

As outlined in section 2.2, PD data types are diverse. For example, they can be reported by binary responses (e.g. presence of disease) or count data (e.g. the number of episode of a particular condition). Logistic models and poisson regression models respectively would be required in these cases. Here, for simplicity, we focus only on a continuous PD response. Let $e_{ij'}$ denote the drug response for subject i measured at j' time point, for $i = 1, \cdots, m$ and $j' = 1, \cdots, n_i'$. ϕ_i is a p'-dimensional vector of individual-specific PD parameter. The sampling distribution is described as

$$e_{ij'}|\phi_i, \theta_i, \sigma_{pk}^2, \sigma_{pd}^2 \sim N(f'(\phi_i; \theta_i; t_{ij'}), \sigma_{pd}^2), \tag{3.16}$$

where $f'(\cdot)$ is the function of individual-specific PD parameters $\boldsymbol{\phi}_i$. In general, the PK model is a compartmental model so one can describe concentrations profile at the "effect compartment". Thus we assume the dynamics depend upon the kinetics only through the "effect compartment" concentrations, $C_e(\cdot)$:

$$f'(\phi_i; \theta_i; t_{ij'}) = h(\phi_i, C_e(\theta_i, t_{ij'}; D_i)) \tag{3.17}$$

A simple PK/PD "link model" can then be identified as,

$$p(e_{ij'}|\phi_i, \theta_i; \sigma_{pd}^2) = N(h(\phi_i; C_e); \sigma_{pd}^2) \tag{3.18}$$

where the function $h(\cdot)$ is given by the classic E_{max} formula described by

$$h = \frac{E_{max} \times C_e}{C_e + C_{50}} = \frac{\phi_{i1} C_e}{C_e + \phi_{i2}} \tag{3.19}$$

where ϕ_{i1} and ϕ_{i2} indicating maximal effect and the concentration where the drug effect reaches maximal for subject i.

Similar as in the PK model, the hierarchical model for the PD response is expressed by

$$p(\phi_i) \sim N_{p'}(Z_i' \mu', \Sigma_{pd}) \tag{3.20}$$

$$\sigma^2_{pd} \sim IG(a', b'); \qquad \mu' \sim N_s(\eta', H'); \qquad \Sigma_{pd} \sim IW(R', \rho') \tag{3.21}$$

Here Z'_i is an $p' \times s$ covariance-effect design matrix, μ' is a vector of s fixed effect parameters, and Σ_{pd} is $p' \times p'$ covariance matrix for PD parameters.

In the following, we denote the matrix of all observed PK data $Y = y_{ij}$, $i = 1, \cdots m$ and $j = 1, \cdots, n_i$, and similarly let E denote all observed PD data. And let P be the set of population parameter of interest, i.e. $\mu, \mu', \Sigma_{pk}, \Sigma_{pd}, \sigma^2_{pk}, \sigma^2_{pd}$.

A typical Bayesian analysis involves estimation of the joint posterior distribution of all unobserved quantities conditional on the observed data. Bayes' theorem allows us to express the posterior distribution as follows,

$$\begin{aligned} p(\theta, \phi, P|Y, E) &= \frac{p(Y, E|\theta, \phi, P)p(\theta, \phi, P)}{p(Y, E)} \\ &\propto p(Y, E|\theta, \phi, P)p(\theta, \phi, P) \\ &= p(Y, E, \theta, \phi|P)p(P) \end{aligned} \tag{3.22}$$

More specifically, $p(Y, E, \theta, \phi|P)$ in equation 3.22 can be written as

$$\begin{aligned} p(Y, E, \theta, \phi|P) = &\Big\{ \prod_{i=1}^m \prod_{j=1}^{n'_i} p(e_{ij'}|\phi_i, \theta_i, \sigma^2_{pd}) \Big\} \times \Big\{ \prod_{i=1}^m \prod_{j=1}^{n_i} p(y_{ij}|\theta_i, \sigma^2_{pk}) \Big\} \\ &\times \Big\{ \prod_{i=1}^m p(\phi_i|\theta_i, \mu', \Sigma') \Big\} \Big\{ \prod_{i=1}^m p(\theta_i|\mu, \Sigma) \Big\} \end{aligned} \tag{3.23}$$

which is the full likelihood function included in the model. The distribution on

$$p(P) = p(\mu)p(\mu')p(\Sigma_{pk})p(\Sigma_{pd})p(\sigma^2_{pk})p(\sigma^2_{pd}). \tag{3.24}$$

The posterior is proportional to the multiplication of 3.23 and 3.24. This is the whole structure of a complete PK/PD model.

We can use MCMC methods to explore the joint posterior distribution of interest. We can infer any interested quantity as long as their realization can be generated in an ergodic Markov chain sequence of draws from full-conditional distributions. Monte Carlo simulation from a Markov chain requires that the the stationary distribution is the target posterior.

There are various algorithms for computing, such as Gibbs sampling, Metropolis-Hastings. The Gibbs sampler is the primary tool to iteratively simulate each quantity from the full conditional distribution. Given proper initial values for each qunantity of interest, one iteration of the Gibbs sampler iterates through sampling across the following densities,

$$\begin{aligned}
&\theta^{(1)} \sim p(\theta|\phi^{(0)}, \mu^{(0)}, \mu'^{(0)}, \sigma_{pk}^{2(0)}, \Sigma_{pk}^{(0)}, \sigma_{pd}^{2(0)}, \Sigma_{pd}^{(0)}, E, Y) \\
&\phi^{(1)} \sim p(\phi|\theta^{(1)}, \mu^{(0)}, \mu'^{(0)}, \sigma_{pk}^{2(0)}, \Sigma_{pk}^{(0)}, \sigma_{pd}^{2(0)}, \Sigma_{pd}^{(0)}, E, Y) \\
&\Sigma_{pk}^{(1)} \sim p(\Sigma|\phi^{(1)}, \theta^{(1)}, \mu^{(0)}, \mu'^{(0)}, \sigma_{pk}^{2(0)}, \sigma_{pd}^{2(0)}, \Sigma_{pd}^{(0)}, E, Y) \\
&\text{etc.}
\end{aligned} \tag{3.25}$$

After T iterations, we have samples $\theta^{(T)}, \phi^{(T)}$ and $P^{(T)}$. When $T \to \infty$,

$$p(\theta^{(T)}, \phi^{(T)}, P^{(T)}) \overset{d}{\sim} p(\theta, \phi, P|Y, E) \tag{3.26}$$

where "$\overset{d}{\sim}$" represents "convergence in distribution". If a Gibbs sampler is not suitable, one can use a MH algorithm, rejection sampling as in Gilks et al. (1992), "slice" sampling as in Neal et al. (1997).

3.2.2 WinBUGS, PKBUGS and ADAPT

There are three popular programs to implement the PK/PD models. The most popular and versatile Bayesian program is WinBUGS. The package can handle complex Bayesian analyses using MCMC methods. Since WinBUG requires a method specification that is suitable for any class of model, it is not optimal for population PK/PD models. Because the specification of population PK/PD models for majority of "real-life" application is complex, such as the complexity of patients's dosing histories, time-varying covariates, censored observations or missing data.

PKBUGS alleviates those difficulties of model specification and thus makes state-of-art MCMC techniques accessible to the analysis of PK-PD models. It is an interface for

the Bayesian statistical software for the analysis of only pharmacokinetic data. The main feature of PKBUGS is to simplify the specification of PK modeling by using dialog boxes and menu commands. It also can recognize the NONMEN data format and a number of standard data items (like the patients's id, time and response). The users can regress the covariates against the desired PK parameters. The structure specifies 28 PK compartment models with the input characteristics, such as bolus/infusion intravenous, first-order with initial lag time. PKBUGS is a customized-oriented hierarchical PK model. See more in Lunn $(1999), (2002)$, Lunn and Aarons $(1997), (1998)$.

ADAPT is another computational modeling platform developed for pharmacokinetics and pharmacodynamics applications. It provides almost all relevant parametric nonlinear mixed-effects modeling algorithms, such as the first-order conditional estimation method, Laplace method, the Monte Carlo parametric Expectation Maximization and MCMC algorithms. It is designed to describe the biological process, such as PK/PD, estimate model parameters and make inference from model prediction. Bauer et al. (2007) used three published data to compare the results from ADAPT with that of NONMEN and WINBUGS. He claimed that ADAPTs performance was very stable, more efficiently with more complex PK/PD models involving a system of differential equations. See more applications in Hong, 2007, Ng CM, $2013, 2010$.

Chapter 4

Bayesian Nonparametric Modeling

The parametric statistics assume that the data come from a type of probability distributions and make inference on the parameters of the distribution. These probability distributions are characterized by a finite number of parameters, i.e. a normal distribution with unknown mean and variance. In contrast, nonparametric statistics avoids assumptions on the probability distribution, for example a classical nonparametric test, the sign test.

In general, parametric methods make more assumptions than non-parametric methods. If those assumptions are not violated, parametric methods can produce accurate and precise estimates. However, if the parametric assumptions are violated, parametric methods can be misleading. For this reason, parametric approaches are not robust. Nonparametric approaches are robust because they generally do not have assumptions on distributions. What's more, when modeling a distribution over data, parametric models use a fixed and finite number of parameters. Thus, they can suffer from over- and under-fitting of the data. It becomes difficult to balance between the complexity of the model and the amount of data available. In this situation, non-parametric Bayes provides an alternative approach to parametric modeling and allows the number of parameters to change when more data are collected. Thus it can avoid under-fitting and over-fitting of the model.

Suppose we have an underlying and unknown distribution which we wish to infer given some observed data. Say, we observe $x_1, x_2, \cdots$, with $x_i \sim F$. In other words, we assume that the observations are independent and identical draws from F. A Bayesian would approach this problem by placing a prior over F, then computing the posterior over F given data. Traditionally these models and priors are chosen from a parametric family. However, restricting distributions to a certain parametric family may limit the scope and type of inference that can be made. The nonparametric Bayesian approach uses a prior with a wide support, i.e. prior distribution can change with more data observed. We can make different types of inference given a large space when posterior computations are tractable.

The Dirichlet process (DP) is one of the most popular Bayesian nonparametric models. Ferguson et al. (1973) first formalized it for general Bayesian statistical modeling. DP can be succinctly described as a distribution over distributions, i.e. each draw from a DP is a distribution itself. It is called a DP because it is characterised by Dirichlet distributed finite dimensional marginal distributions. This is similar to a Gaussian process, where the finite dimensional marginal distribution are Gaussian distributions. But the distributions drawn from a DP are discrete, and they can not described using a finite number of parameters. Since the random draw from DP is a discrete distribution, it is often used for clustering the population into heterogeneous subpopulation. In this chapter, we introduce the DP and DP mixture model with MCMC techniques in the Bayesian framework. In section 4.1, we present the Dirichlet Process and its properties. Section 4.2 discusses DP mixture model. In section 4.3, we discuss a modern MCMC sampling algorithm of Bayesian inference in DP mixture model.

4.1 The Dirichlet Process

The Dirichlet Process (Ferguson, 1973) is one of the most widely used Bayesian nonparametric models for modeling unknown random distributions in Bayesian statistics. It has been applied to a wide range of problems, such as variable selection in genetics (Kim et al. 2006), linguistics (Teh, 2006b), psychology (Navarro et al., 2006, image segmentation (Sudderth and Jordan, 2009) as well as in the neurosciences (Jbabdi et al., 2009). In this section, we will review the popular representations of DP in Bayesian framework.

4.1.1 Definitions

There are many ways to define a Dirichlet process. Before we proceed to a formal definition, let us see an intuitive definition first. Consider a Bayesian mixture model consisting of K components,

$$x_i|z_i, \theta_k^\star \sim F(\theta_k^\star) \tag{4.1}$$

with

$$z_i|\pi_\alpha \sim \text{Multinomial}(\pi_\alpha); \quad \pi_\alpha \sim Dir(\alpha/K, \cdots, \alpha/K); \quad \theta_k^\star|H \sim H \tag{4.2}$$

where π_α denotes the vector of mixing proportion, α is the concentration parameter in Dirichlet distribution, and H is a distribution over $\theta_k^\star$. F is a parametric distribution indexed by θ_i. Then $x_i \sim \sum_{k=1}^K \pi_k F(\theta_k^\star)$ which defines a finite mixture model. Here x_is are random draws from a component whose distribution parameterized by $\theta_k^\star$. Each observation x_i has a class label z_i indicating which component x_i is from. z_i follows a multinomial distribution with a parameter vector from a Dirichlet distribution. When $K \to \infty$, we obtain an infinite mixture model.

Ferguson (1973) first developed DP considering its finite dimensional distributions. For a random distribution $G \sim DP(\alpha, H)$, its marginal distributions must be Dirichlet

distributed. A formal definition is that we say G is a Dirichlet process with base distribution H and concentration parameter α, written $G \sim DP(\alpha, H)$, if

$$(G(A_1), G(A_2), \cdots, G(A_r)) \sim Dir(\alpha H(A_1), \alpha H(A_2), \cdots, \alpha H(A_r)) \tag{4.3}$$

for any finite measurable partition $A_1, \cdots, A_r$ of Ω. Here $Dir(\gamma_1, \cdots, \gamma_k)$ denotes the k-dimensional Dirichlet distribution with parameters $\gamma_1, \cdots, \gamma_k$.

The parameters H and α play an intuitive role in the definition of DP. The base distribution is the mean of the DP. For a measurable set $A \in \Omega$, we have $E[G(A)] = H(A)$. On the other hand, the concentration parameter can be understood as a precision parameter: $Var[G(A)] = \frac{\alpha H(A)(\alpha - \alpha H(A))}{\alpha^2(\alpha+1)} = \frac{H(A)(1-H(A))}{\alpha+1}$. The larger α is, the smaller the variance, and then DP will concentrate more of its mass around the mean. α is also called strength parameter referring to the strength of the prior when using DP as a prior in a Bayesian nonparametric model. See more explanation in posterior distribution of DP next.

4.1.2 Properties of Dirichlet Process

The DP has several well-known representations and properties. Here we give a brief summary.

Polya Urn Scheme and De Finetti's Theorem

Polya Urn scheme provides a visualization of DP (Blackwell and Macqueen, 1973). More importantly, it can be used to prove the existence of DP. Here we introduce the Polya Urn scheme first and discuss the existence of DP. Suppose each value in the space Ω has a unique color. In the beginning, there are no balls in the urn. First pick a color from H, i.e. $\theta \sim H$, paint a ball with that color and drop it into urn. Second, either pick a ball from the urn, paint a new ball with the same color, then drop both balls into the urn; or draw a new

color from H, paint a ball with that color, and drop it into the urn. In the subsequent steps, for the $(n+1)$th time, draw a color from H with probability $\frac{\alpha}{\alpha+n}$, paint a new ball with that color and drop the ball into the urn; or choose a ball from the urn with probability $\frac{n}{\alpha+n}$, paint a new ball with the same color and drop two balls back into urn. The color of an infinite sequence of draws follows a DP with strength parameter α and base distribution H.

The distribution over the color of ball represents the DP. The distribution of new ball color depends the distribution of existing color and the base measure H. The sequence of ball colors has an exchangeability property. Consider a finite sequence of ball colors, $\theta_1, \ldots, \theta_n$, then we can define a distribution over the first n ball's color as,

$$P(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} P(\theta_i | \theta_1, \ldots, \theta_{n-1}) \tag{4.4}$$

The right-hand side in 4.4 is the same as any sequences of n elements from H. For example, given any permutation σ of $\{1, 2, \ldots, n\}$, it is not difficult to show,

$$P(\theta_1, \ldots, \theta_n) = P(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n-1)}). \tag{4.5}$$

The joint density of a sequence of draws are equal to the joint density of any permutation of the sequence. This can be extended to the infinite sequence of DP. De Finetti's theorem states that for any infinitely exchangeable sequence, $\theta_1, \theta_2, \ldots$, there is a random distribution such that the sequence of independently identical draws are from it,

$$P(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} G(\theta_i) dP(G) \tag{4.6}$$

In this setting, $G \sim DP(\alpha, H)$ is the mixing De Finetti's measure.

Posterior distribution

Here we explore the posterior distribution of a DP. Let $G \sim DP(\alpha, H)$. Since G is a random distribution, we can draw samples from G. Let $\theta_1, \theta_2, \cdots, \theta_n$ be a random sample

from G. We are interested in the posterior distribution of G given θ_i. To examine it, let $A_1, \ldots, A_r$ be a finite partition of parameter space Ω. Let n_k denote the number of θs in A_k, $n_k = \#\{i : \theta_i \in A_k\}$. $G \sim DP(\alpha, H)$, by definition of DP,

$$\begin{aligned}(G(A_1), G(A_2), \cdots, G(A_r) \quad &| \quad \theta_1, \theta_2, \cdots, \theta_n) \\ &\sim Dir(\alpha H(A_1) + n_1, \alpha H(A_2) + n_2, \cdots, \alpha H(A_r) + n_r)\end{aligned}$$

Since it is true for any partition of Ω, the posterior of distribution of G must be a DP too. Now we can figure out the updated parameters for the posterior. The new concentration parameter is equal to $\alpha H(A_1) + n_1 + \cdots + \alpha H(A_r) + n_r = \alpha + n$. The new base distribution is a mixture of the empirical cumulative distribution function of θ's and n, written as $\frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$, where $\delta_{\cdot}$ is the point mass at θ_i. The posterior can be rewritten as

$$G|\theta_1, \theta_2, \cdots, \theta_n \sim DP(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}) \tag{4.7}$$

The posterior base distribution is a weighted average between the prior base distribution H and the empirical distribution $\sum_{i=1}^{n} \delta_{\theta_i}$. The weight associated to n has weight proportional to α and the empirical distribution has weight proportional to the number of observations n. From this representation, it follows that α can be interpreted as a strength parameter referring the strength of the prior.

Predictive Distribution

Now we explore the predictive distribution of draws from a DP. Since $G|\theta_1, \cdots, \theta_n$ is a DP, we have

$$\begin{aligned} p(\theta_{n+1} \in A|\theta_1, \cdots, \theta_n) &= E(G(A)|\theta_1, \cdots, \theta_n) \\ &= \frac{\alpha}{\alpha + n} H(A) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}, \end{aligned}$$

for any measurable set $A \in \Omega$, where the last step is the posterior base distribution of G given the first n observations. Marginalizing out G, we obtain the predictive distribution,

$$\theta_{n+1}|\theta_1, \cdots, \theta_n \sim \frac{\alpha}{\alpha+n}H + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\theta_i} \tag{4.8}$$

Given the posterior base measure is also the predictive distribution of θ_{n+1}. When $\alpha \to 0$, the prior becomes non-informative, that is the predictive distribution is given by the empirical distribution. When the number of observations is large, $n >> \alpha$, the predictive distribution is dominated by the base distribution which is also close to the underlying distribution. This indicates the property of the DP: the posterior of DP converges to the true underlying distribution. We can also see that DP is discrete with sum of probability one. Equation 4.8 helps us to understand that each draw from DP is a distribution. The first draw θ_1 can be defined as a point mass at θ_i. The distribution of the second draw is $\theta_2|\theta_1 \sim \frac{\alpha}{\alpha+1}H + \frac{1}{\alpha+1}\theta_1$ and so on. The DP can be succinctly described as a distribution over distributions.

The predictive distribution can often be characterized as a species sampling(SS) allocation rule. More formally, a SS sequence is a sequence of random variables $\theta_1, \theta_2, \cdots$, characterized by the predictive probability function,

$$P\{\theta_{n+1} \in \cdot|\theta_1, \cdots, \theta_n\} = \sum_{j=1}^{n} q_{n,j}\delta_{\theta_j}(\cdot) + q_{n,n+1}H \tag{4.9}$$

Here $q_{n,j} = \frac{1}{\alpha+n}$ and $q_{n,n+1} = \frac{\alpha}{\alpha+n}$ are non-negative values with $\sum_{j=1}^{n+1} q_{n,j} = 1$. H is a non-atomic probability measure (Pitman, 1996b). Considering the unique values of θ_j, the equation 4.9 can be rewritten as

$$P\{\theta_{n+1} \in \cdot|\theta_1, \cdots, \theta_n\} = \sum_{j=1}^{K_n} q_j^\star\delta_{\theta_j^\star}(\cdot) + q_{K_n+1}^\star H, \tag{4.10}$$

where K_n is the number of unique values, say $(\theta_1^\star, \cdots, \theta_{K_n}^\star)$ are the unique elements in the sample of $(\theta_1, \cdots, \theta_n)$. $q_j^\star$s are the normalizing constants. Here $q_j^\star$ only depends on the

frequency of θ_j occurring. From 4.10, we know that the value $\theta_k^\star$ will be drawn θ_{n+1} with probability proportion to n_k, the number of times the value has already been observed. In the first n draws, the larger n_k is, the higher the probability that the new draw θ_{n+1} equals to $\theta_k^\star$. This is a rich-gets-richer phenomenon, where large clusters (a set of θ_i with same $\theta_k^\star$) grows larger fast. It also shows the clustering property of the DP by looking at partitions induced by the clustering. The unique values of $\theta_1, \ldots, \theta_n$ induce a partitioning of the set $[n] = \{1, \ldots, n\}$ into clusters such that in each cluster, θ_i's take on the same values $\theta_k^\star$. Since $\theta_1, \ldots, \theta_n$ are random, this induces a random partition of $[n]$. Larger n_k and lower value of α imply a tighter clustering.

The Chinese Restaurant Process (CRP) is another popular metaphor used to interpret DP. Imagine a Chinese restaurant with an infinite number of tables, and a sequence of customers are waiting for entering. The first customer enters and sits at the first table. The second customer enters and sits either with the first customer at first table with probability $\frac{1}{1+\alpha}$, or sits at a new table with probability $\frac{\alpha}{1+\alpha}$. In the subsequent steps, n^{th} customer comes, he/she can sit at occupied table K with the probability $\frac{k}{n-1+\alpha}$, where k represents the number of previous customers already sitting at table K, or at a new table with probability $\frac{\alpha}{n-1+\alpha}$. Identifying customers with integers $1, 2, \ldots$ and tables as clusters, n customers define a partition of $[n]$. CRP defines a distribution over partitions of $[k]$,$k = 1, 2, \cdots, K$. The value of k is the number of customers. The distribution over tables is a DP.

Stick-breaking Construction

Another representation of the Dirichlet process is provided by the stick-breaking construction (Sethuraman, 1994, Pitman, 1996). This process can be used to provide a constructive algorithm for generating a DP. The distribution of DP is given by the density of a weighted

sum of point masses.

$$G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^\star}$$

$$\text{with} \quad \pi_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i); \quad \beta_k \sim Beta(1,\alpha); \quad \theta_k^\star \sim H \tag{4.11}$$

The construction of π can be understood as starting with a stick of length 1. Break it at β_1, assigning π_1 to be the length of stick we just broke off, assign π_i on a draw θ_i from H. Break the remaining portion of the stick at β_2 with the breaking off length π_2, assign π_2 as a corresponding weight for the second draw θ_2 from H. Recursively break the remaining portion to obtain π_3 and so forth. The stick-breaking distribution over π is written as $\pi \sim GEM(\alpha)$, where the letters stand for Griffiths, Engen and McCloskey.

This representation is the most versatile definition of the Dirichlet process. It has been explored to generate efficient alternative MCMC algorithms. It is also the basis of the definition of the generalizations that allow dependence across a collection of distributions, i.e. dependent Dirichlet process (MacEachern, 2000, Griffin and Steel (2006)).

4.2 Dirichlet Process Mixture model

Dirichlet process mixture models go back to Antoniak (1974) and Ferguson (1973). They have been developed by Escobar and West (1995), MacEachern and Muller (1998). DPM is one of the most classic models in nonparametric Bayesian. It defines a mixture model with countably infinitely components and can be used in density estimation or clustering while the number of components is a priori unknown. DPM have become increasingly popular for modeling when traditional parametric models impose unreasonably constraints on the distribution. Examples of applications includes empirical Bayes problems (Escobar, 1994), nonparametric regression (Muller, Erkanli, and West (1996)), density estimation

(Escobar and West 1995), hierarchical modeling (MacEachern 1994, West, Muller, and Escobar 1994).

The core of the DPM model can basically be thought of as a simple Bayesian model. Data $y_1, \cdots, y_n$ are independently draws from unknown distribution F. The prior is $\theta_i \sim G$. Here we add uncertainty about the prior distribution G, instead of using a distribution from the exponential family.

$$y_i|\theta_i \sim F(\theta_i), \qquad \theta_i|G \sim G, \qquad G \sim DP(\alpha, H) \tag{4.12}$$

where G is a random distribution from DP with base measure H and concentration parameter α, $F(\cdot)$ is an unknown distribution with parameter θ_i. The general applications of DPM typically allow the introduction of subject-specific covariates. And the more complex models also introduce distributions on the hyper-parameters on F, α and H.

Before we move to estimate the DPM and discuss MCMC sampling algorithms, we need to briefly prove that an equivalent limiting process of a Dirichlet process mixture model. This limiting process can improve the sampling efficiency. We start with a finite mixture model with k components. The likelihood function can be written as

$$p(y_1, \cdots, y_n) = \prod_{j=1}^{k} \pi_j \prod_{i \in j}^{n} p(y_i|\theta_j) \tag{4.13}$$

where π_j is the mixing proportion for component j with $\sum \pi_j = 1$. $p(\cdot|\theta_j)$ is the probability density function for component j with parameter θ_j. Here we first consider the model for a fixed k components, then explore more properties as $k \to \infty$.

The mixing proportions π_j, are given a symmetric Dirichlet prior with parameter $\alpha/k, \cdots, \alpha/k$:

$$p(\pi_1, \cdots, \pi_k|\alpha) \sim D(\alpha/k, \cdots, \alpha/k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^{k} \pi_j^{\alpha/k-1} \tag{4.14}$$

Here we introduce an indicator variable c_i to "label" which component generates x_i. c_is takes on values $1, \cdots, k$. The joint distribution of the c_is is multinomial with parameter $\boldsymbol{\pi} = \pi_1, \cdots, \pi_k$ expressed,

$$p(c_1, \cdots, c_n|\boldsymbol{\pi}) \propto \prod_{j=1}^{k} \pi_j^{n_j}, \qquad n_j = \sum_{i=1}^{n} \delta_{c_i}(j) \tag{4.15}$$

Integrating out the mixing proportions $\boldsymbol{\pi}$, the joint density of $c_1, \cdots, c_n$,

$$\begin{aligned} p(c_1, \cdots, c_k) &= \int p(c_1, \cdots, c_k|\pi_1, \cdots, \pi_k) d\pi_1, \cdots, d\pi_k \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \int \prod_{j=1}^{k} \pi_j^{n_j+\alpha/k-1} d\pi_j \\ &= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=1}^{k} \frac{\Gamma(n_j+\alpha/k)}{\Gamma(\alpha/k)} \end{aligned} \tag{4.16}$$

In order to use Gibbs sampling to update c_i in MCMC simulation, we need investigate the conditional prior c_i given the rest of labels, denoted as $c_i|c_{-i}$. Assuming $c_i = c$,

$$\begin{aligned} p(c_i = c|c_{-i}) &= p(c_1, \cdots, c_i = c, \cdots, c_n)/p(c_1, \cdots, c_{i-1}, c_{i+1}, \cdots, c_n) \\ &= \frac{\int p_{c_1}, \cdots, p_{c_i=c}, \cdots, p_{c_n} \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^{-k}} p_1^{(\alpha/k)-1}, \cdots, p_k^{(\alpha/k)-1} d\boldsymbol{p}}{\int p_{c_1}, \cdots, p_{c_{i-1}}, p_{c_{i+1}}, \cdots, p_{c_n} \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^{-k}} p_1^{(\alpha/k)-1}, \cdots, p_k^{(\alpha/k)-1} d\boldsymbol{p}} \\ &= \frac{n_{-i,j}}{n-1+\alpha} + \frac{\alpha/k}{n-1+\alpha} \end{aligned} \tag{4.17}$$

In summary, the above finite mixture model can be expressed hierarchically as follows,

$$y_i|c_i, \phi_{c_i} \sim f(y_i|\phi_{c_i}) \tag{4.18}$$

with

$$c|\pi_{1:k} \sim Dis(\pi_1, \ldots, \pi_k); \quad \phi_{c_i} \sim H(\theta_i); \quad \pi_1, \ldots, \pi_k \sim Dir(\alpha/k, \ldots, \alpha/k) \tag{4.19}$$

where Dis denotes the discrete distribution. Dir represents the Dirichlet process. When $k \to \infty$, the conditional probabilities defining the prior for the c_i reach the limits,

$$p(c_i = c|c_{-i}) \quad \to \quad \frac{n_{-i,c}}{n-1+\alpha} \tag{4.20}$$

If $c_i \neq c_j$ for all $j \neq i$,

$$p(c_i|c_{-i}) \quad \to \quad \frac{n_{-i,c}}{n-1+\alpha} \tag{4.21}$$

This limiting process is equivalent to Dirichlet process mixture model if we consider $\theta_i = \phi_{c_i}$.

4.3 Inference samplings in DPM model

Use of DPM models has become computationally feasible with the development of Markov chain methods for sampling from the posterior distribution. Methods based on Gibbs sampling can be easily to implement in the model based on conjugate prior. However, in non-conjugate case, it is difficult to carry out because of the difficulty of solving numerical integral. West, Muller and Escobar (1994) used a Monte Carlo approximation to the integral. MacEchern and Muller (1998) developed an exact method for non-conjugate priors. They used a mapping from a set of auxiliary parameters to the set of parameters currently in use. Neal (2000) presented a comprehensive survey of MCMC sampling in DP mixture models for both conjugate and non-conjugate priors. In our real application, the prior is not conjugate but we developed a modern sampling algorithm that is suitable for our special population PK/PD data. Here we first introduce a simple sampling algorithm for a conjugate prior, and then move to a complex algorithm for non-conjugate case.

Since a draw from a DP is discrete random distribution. Given a sequence of random draws $\theta_1, \cdots, \theta_n$, the prior distribution of the θ_i can be expressed as

$$\theta_i|\theta_{-i} \sim \frac{1}{n-1+\alpha}\sum_{j\neq i}\delta_{\theta_j} + \frac{\alpha}{n-1+\alpha}H \tag{4.22}$$

where δ_θ is the point mass at θ_j.

Given the likelihood and prior, the posterior distribution $\theta_i|\theta_{-i}$, the posterior distribution $\theta_i|\theta_{-i}, y_i$ is,

$$\begin{aligned}\theta_i|\theta_{-i}, y_i &\sim \frac{1}{i-1+\alpha}\sum_{j\neq i}p(y_i|\theta_j)\delta_{\theta_j} + \frac{\alpha}{i-1+\alpha}p(y_i|\theta_i)H(\theta) \\ &= \frac{1}{i-1+\alpha}\sum_{j\neq i}p(y_i|\theta_j)\delta_{\theta_j} + \frac{\alpha}{i-1+\alpha}p(\theta|y_i, H)\int p(y_i|\theta_i)dH(\theta)\end{aligned} \tag{4.23}$$

where $p(\theta_i|y_i, H)$ is the posterior distribution of θ based on an observation y_i and H. If H is conjugate prior for the likelihood F, we can use Gibbs sampling to repeatedly draw the new values from 4.23 and then make inference about θ_i. Because it is not difficult to compute the integral $\int p(y_i|\theta_i)dH(\theta)$ and sampling from $p(\theta_i|y_i, H)$. Escobar (1994) and Escobar and West (1995) presented a simple algorithm based on equation 4.23. The sampling is not efficient because of two reasons. First, when we update θ_i, it goes through the low-probability states before reaching to the highest probability state. It becomes slower if we have more states in practice. Second, we can only renew θ_i no more once at each iteration.

Bush and MacEachern (1996) and West, Muller and Escobar (1994) improved this algorithm by borrowing the strength of the limiting process. The basic idea is to assign a cluster label c_i for y_i, update c_i using the required probability for $i = 1, \cdots, n$. Some observations are assigned into several groups. In each group, we may have one or more observations and then we can sample $\theta_{c_i=c}$ given $y_i \in c$. This algorithm allows us to change θ_i at least twice and it may pass through the inter-mediate state fast. Here we

presents more details about the implementation. The permanent chain includes $\{c_i, \theta_i\}$. We take updating θ_i as an example. To keep the simplicity of the notation, let c_{-i} denote the all c_js , $j \neq i$, similarly for θ_{-i} too. k^- denotes the number of distinct values of θ_{-i} or c_{-i}. If $c_i = c_j$ for some $j \neq i$s, the required conditional probability is,

$$p(c_i = c|c_{-i}, \theta_i, y_i) \propto \frac{n_{-i,c}}{n-1+\alpha} p(y_i|\theta_c) \tag{4.24}$$

where $n_{-i,c}$ denotes the number of $c_{-i} = c$. If $c_i \neq c_j$ for all c_js, $j \neq i$, the conditional probability for c_i is

$$p(c_i \neq c_j|c_i, \theta_{-i}) \propto \frac{\alpha}{n-1+\alpha} \int p(y_i, \theta) dH(\theta) \tag{4.25}$$

Using the equation 4.24 and 4.25, we can update c_i for $i = 1, \cdots, n$. Suppose that we have K groups in total after updating the cluster label. In each group, we may have several observations. Therefore, we can sample θ_k using Metropolis-Hastings algorithm.

If H is a non-conjugate prior on F, it is difficult to compute the integral, especially in high-dimensional cases. In order to avoid the integral computation and create a chain mixing fast, we implement an adaptive MCMC algorithm, Gibbs sampling with auxiliary parameters, in our real application.

The basic idea of the modern MCMC algorithm is that we are interested in sampling θ_i from $p(\theta_i|\theta_{-i}, y_i)$. However, we can not sample it directly because of the integral computation $\int p(y_i, \theta) dH(\theta)$. Instead we introduce several auxiliary parameters from H, draw θ_i from the conditional distribution $p(\theta_i|\theta_{-i}, \theta_{auxiliary}, y_i)$ and discard $\theta_{auxiliary}$ during the Markov chain simulation. We know that the $p(\theta_i|\theta_{-i}, \theta_{auxiliary}, y_i)$ is the marginal distribution $p(\theta_i, \theta_{auxiliary})$. It is easy to see that this update for θ_i will leave $\theta_i|\theta_{-i}$ invariant. In the end, keep θ_i and discard $\theta_{auxiliary}$s. The sampling process can be expressed as, first, draw a new value of $\theta_{auxiliary}$ from H, given $p(\theta_i|\theta_{-i})$ from the conditional distribution; second, update θ_i given the new value of $\theta_{auxiliary}$ from the first step; third, discard $\theta_{auxiliary}$ and keep θ_i only. Note that the values of $\theta_{auxiliary}$ are introduced temporarily and they are not

associated with any observations. This algorithm is called Gibbs sampling with auxiliary components (Neal et al. (2000)).

Next we show the general steps to implement the algorithm. The permanent state of the Markov chain has c_i and θ_i. We take sample c_i and θ_i as an example. First we introduce a number of temporary auxiliary components, denoted $\theta_{k^-+1}, \cdots, \theta_{k^-+m}$. Here k^- denotes the number of distinct value among c_j or $\theta_j, j \neq i$.

The first step is to update c_i based on the conditional distribution c_i given all other parameters associated with the observations and the auxiliary parameters. If $c_i = c_j = c$ for some js, which means $c_i \in \{1, 2, \cdots, k^-\}$, independently draw m auxiliary components from H, the conditional distribution of c_i is,

$$c_i | c_{-i}, \theta_{1:k^-}, \theta_{k^-+1}, \cdots, \theta_{k^-+m}, y_i \propto \frac{n_{-i,c}}{n-1+\alpha} p(y_i, \theta_c) \tag{4.26}$$

If $c_i \neq c_j$ for all $j \neq i$, assign the class label $c_i = k^- + 1$, and let $\theta^{k^-+1} = \theta_i$. We draw the rest of $m-1$ auxiliary components from H. The corresponding distribution is,

$$p(c_i = k^- + 1 | c_{-i}, \theta_{1:k^-}, \theta_{k^-+1}, \cdots, \theta_{k^-+m}, y_i) \propto \frac{\alpha/m}{n-1+\alpha} p(y_i, \theta_c) \tag{4.27}$$

where $c \in (k^-+1, \cdots, k^-+m)$. The equation 4.27 shows that θ_i has an equal probability to be one of auxiliary parameters.

After updating c_i for $i = 1, \cdots, n$, some observations share the same class label which indicates that they have the same parameter values in a group. We use $\theta_1^\star, \cdots, \theta_k^\star$ denote the distinct values of $(\theta_1, \cdots, \theta_n)$. Sampling $\theta_c^\star$ given all the observations $y_i, i \in c$ is straightforward. The target distribution equals the multiplication of the likelihood function and prior,

$$p(\theta_k^\star | y_{i \in k}, c_i = k) = \prod_{i \in k} p(y_i | \theta_k^\star) \pi(\theta) \tag{4.28}$$

To simulate from the target distribution $p(\theta_k | \cdot)$, we start with a starting values $\theta^{(0)}$. At

iteration T, we draw a proposal $\theta^{\star}$ from a known proposal distribution $q(\theta)$. We accept

$$\theta^{(T)} = \begin{cases} \theta^{\star}, & \text{with probability } \alpha \\ \theta^{T-1}, & \text{with probability } (1-\alpha) \end{cases}$$

with α defined:

$$\alpha = \min\Big\{ \frac{p(\theta^{\star}|y)}{p(\theta^{(T-1)}|y)} \frac{q(\theta^{(T-1)}|\theta^{\star})}{q(\theta^{\star}|\theta^{(T-1)})}, \quad 1 \Big\} \tag{4.29}$$

To implement the M-H algorithm, it is important to choose a proposal distribution such that the chain "mixes well", i.e. adequately explores the posterior distribution. The popular choice is the multivariate normal distribution, Bennett Racine-Poon (1996), Carlin and Louis (1996), and Huang (2004). However, an important issue regarding the multivariate normal proposal distribution is the dispersion of the proposal density. If the variance of the proposed density is too large, the large proportion of proposed candidates will be rejected, and the Markov chain will waste many repeats and result in inefficiency of the algorithm. On the other hand, if the variance of the proposed density is too small, the chain will have a high acceptance rate but will move only in a small parameter space, leading to inefficiency too. Tuning of associated parameters such as proposal variances is crucial to achieve efficient mixing, but can also be very difficult.

In our MCMC implementation, we consider an automated tuning of MCMC algorithm Harrio et al. (2001). Adaptive MCMC algorithms can automatically "learn" better parameter values of Markov chain Monte Carlo algorithms while they run because the proposal variance is the variance of the empirical estimates. We begin with a d-dimensional target distribution $\pi(\cdot)$. We perform a Metropolis algorithm with a proposal distribution given at iteration n by $Q_n(x,) = N(\theta, (0.1)^2 I_d/d)$ for $n \leq 2d$, while for $n > 2d$,

$$Q_n(\theta, \cdot) = (1-\beta)N(\theta, 2.38^2 \Sigma_n/d) + \beta N(\theta, 0.1^2 I_d/d), \tag{4.30}$$

where Σ_n is the current empirical estimate of the covariance structure of the target distribution based on the run so far, and where β is a small positive constant i.e. $\beta = 0.05$, d is

the dimension of θ, and Σ_0 is a fixed non-singular matrix, i.e. $\frac{0.1^2 I_d}{d}$. It is also necessary to use some alternative fixed proposal distribution for the first few iterations when the empirical covariance Σ_n is not yet well-defined. Roberts and Rosenthal (2001) proved that the proposal $N(\theta, 2.38^2\Sigma_n/d)$ is optimal in a particular high-dimensional context. They also demonstrate that this adaptive Metropolis algorithm will indeed "learn" the target covariance matrix and approach an optimal algorithm, even in very high dimension. It may takes many iterations before the adaption significantly improve the algorithm, in the end it will converge enormously faster than a non-adapted random walk Metropolis algorithm. We refer the desired readers to see more details and proof in Harrio et al. (2001) and Roberts and Rosenthal's multiple publications about adaptive Metropolis, Yang (2007), Saksman and Vihola (2008), Bai et al. (2008), Atchade and Fort (2008), Craiu et al. 2008 Bai et al. (2009).

We end this chapter by introducing a sampling step we implemented in our application. After updating c_i for $i = 1, \cdots, n$, assuming that we have K distinct estimates $\theta_1, \cdots, \theta_K$, we consider that it is a random sample from a population distribution $N(\boldsymbol{\mu}, \Sigma)$. To keep the notions clearly, we present the whole structure as follows,

$$\begin{aligned} \boldsymbol{\theta_1, \cdots, \theta_K}|\boldsymbol{\mu}, \Sigma &\sim N(\boldsymbol{\mu}, \Sigma) \\ \boldsymbol{\mu} &\sim N(\boldsymbol{\mu}_0, \Sigma_0) \\ \Sigma &\sim IW(v_0, \Sigma_{iw}^{-1}) \end{aligned} \tag{4.31}$$

where IW represent the inverse-Wishart distribution with mean $\frac{\Sigma_{iw}^{-1}}{v_0 - d - 1}$. The values for $\boldsymbol{\mu}_0, \Sigma_0, v_0$ and Σ_{iw}^{-1} are known.

The posterior distribution of $\mu|\theta_1, \cdots, \theta_k, \Sigma$ is expressed as follows by multiplying the joint likelihood function $\theta_1, \cdots, \theta_K|\mu, \Sigma$, and prior,

$$\begin{aligned} p(\boldsymbol{\mu}|\boldsymbol{\theta_1, \cdots, \theta_K, \Sigma}) &\propto e^{\{-\frac{1}{2}\boldsymbol{\mu^T}\Sigma_0^{-1}\boldsymbol{\mu} + \boldsymbol{\mu^T}\Sigma_0^{-1}\boldsymbol{\mu}_0\}} e^{\{-\frac{1}{2}\boldsymbol{\mu^T K \Sigma^{-1}}\boldsymbol{\mu} + \boldsymbol{\mu^T K \Sigma^{-1}}\bar{\boldsymbol{\theta}}\}} \\ &= e^{\{-\frac{1}{2}\boldsymbol{\mu^T A_n}\boldsymbol{\mu} + \boldsymbol{\mu^T b_n}\}} \end{aligned} \tag{4.32}$$

Equation 4.32 implies that the conditional distribution of $\boldsymbol{\mu}$ must be a multivariate normal distribution with covariance $\boldsymbol{A_n^{-1}} = (\Sigma_0^{-1} + K\Sigma^{-1})^{-1}$ and mean $\boldsymbol{A_n^{-1}b_n} = (\Sigma_0^{-1} + K\Sigma^{-1})^{-1}(\Sigma_0^{-1}\boldsymbol{\mu_0} + K\Sigma^{-1}\bar{\boldsymbol{\theta}})$.

Similarly, given the prior for Σ, the full conditional distribution of Σ given $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K$ and $\boldsymbol{\mu}$ can be written as, The conditional distribution of Σ,

$$\begin{aligned} p(\Sigma|\boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_K}, \boldsymbol{\mu}) &\propto |\Sigma|^{-(v+p+1)/2} e^{-tr(v_0 S_0 \Sigma^{-1})/2} \times |\Sigma|^{-k/2} e^{-tr(S_\theta \Sigma^{-1})/2} \\ &= |\Sigma|^{-(v+k+p+1)/2} e^{-tr(S_0+S_\theta)\Sigma^{-1}/2} \end{aligned} \tag{4.33}$$

Thus we have $(\Sigma|\cdot) \sim W^{-1}(v_0 + k, [S_0 + S_\theta]^{-1})$ with $S_{\boldsymbol{\theta}} = \sum_{i=1}^{k}(\boldsymbol{\theta}_i - \boldsymbol{\mu})(\boldsymbol{\theta}_i - \boldsymbol{\mu})^T$. The result is intuitive. We can think of $v_0 + k$ as the "posterior sample size", the sum of the prior sample size v_0 and the sample size of the data. Similarly, $v_0 S_0 + S_{\boldsymbol{\theta}}$ can be thought of as the "prior" residual sum of squares plus the residual sum of squares from the data.

We can use these full conditional distributions to construct a Gibbs sampler which provides a MCMC approximation to the joint posterior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_k})$. Given a starting value $\boldsymbol{\Sigma^{(0)}}$, the Gibbs sampler generates $\{\boldsymbol{\theta^{s+1}}, \boldsymbol{\Sigma^{s+1}}\}$ from $\{\boldsymbol{\theta^s}, \boldsymbol{\Sigma^s}\}$, first from the equation 4.32 for $\boldsymbol{\mu^{s+1}}$ and then for Σ^{s+1} from the equation 4.33.

Repeatedly sampling to obtain $\{(\mu^{(0)}, \Sigma^{(0)}), \cdots, (\mu^{(T)}, \Sigma^{(T)})\}$. We are interested in the posterior means denoted as $\bar{\theta}_{post} = \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}$ and $\bar{\Sigma}_{post} = \frac{1}{T}\sum_{t=1}^{T}\Sigma^{(t)}$. In our application, we do update μ and Σ given $\theta_1, \cdots, \theta_K$ at each iteration for both PK and PD.

Chapter 5

Nonparametric Bayes model for a population PK/PD model

In this chapter, we will present our proposed approach for a population semi-mechanistic PK-PD model. In section 5.1, we specify our nonparametric Bayes model. Section 5.2 discusses Euler's approximation. In section 5.3, we will discuss the prediction inference. In section 5.4, we will present a simulation data and a real data example.

5.1 Nonparameteric Bayesian Hierarchical Model

We consider a simple version of the model here. For the PK model, we use a two-compartmental model to describe PK response measurement, i.e. unbound plasma concentration. We assume that the measurement variation represents inter-subject variability and an error term. Let y_{ij}^{PK} denote the observed j^{th} measurement of the response in the PK model for subject i at t_{ij} for $i = 1, \cdots, n$, $j = 1, \cdots, n_i$. The value of n is the number

of subjects. The PK model can be described as,

$$y_{ij}^{PK}|\theta_i^{PK} = f_i^{PK}(\theta_i^{PK}, t_{ij}) + \epsilon_{ij} \tag{5.1}$$

where $f_i(\theta_i^{PK}, t_{ij})$ is a function for predicting the jth response in subject i, θ_i^{PK} is a vector of PK parameters. The function f is related to a system of ordinary differential equations, which "represents" what happens to the drug in the body, see Figure $2.1(c)$. Mathematically, we use a differential equation to describes the rate of dose change in each compartment, written as,

$$\begin{aligned}
\frac{dx_c(t)}{dt} &= -(\frac{CL}{V^c} + \frac{CL^d}{V^c})x_c(t) + \frac{CL^d}{V^p}x_p(t) + r(t) \\
\frac{dx_p(t)}{dt} &= \frac{CL}{V^c}x_c(t) - \frac{CL^d}{V^p}x_p(t) \\
C(t) &= \frac{x_c(t)}{V^c}
\end{aligned}$$

where x_c and x_p denote the amount of drug in the central and peripheral compartment separately, $r(t)$ is the infusion rate which is different across patients. CL is the system clearance, V^c and V^p are volumes of the central and peripheral compartment respectively, CL^d is the inter-compartment clearance. The PK parameter is $(CL_i, V_i^c, V_i^p, CL_i^d)$ for subject i. $C(t)$ is the time course of unbound plasma concentration.

For PD part, the observed response is the absolute neutrophil count (ANC) for the cancer patients. Neutrophils are key components in the system of defense against infection. An absence or scarcity of neutrophils (a condition called neutropenia) makes a person vulnerable to infection. After chemotherapy, radiation, the ANC is usually depressed and then slowly rises, reflecting the fact that the bone marrow is recovering and new blood cells are beginning to grow and mature. We use a physiology-based PD model to characterize ANC. The new cells generated in the bone marrow take time to be present in the circulation system. We use a five compartmental model to represent the whole process. We prefer this structure model because of three followed reasons. First, it can separate the system-related parameters which are common across patients, and a drug-related parameter. Second, it

includes three transits compartments which allow us to predict the time delay between the drug effect site and observing site. The third reason is that we can use the full time course of drug concentrations as a special covariate in the PD model. The whole structure consists five compartments, a stem or progenitor compartment, denoted as $Prol$, three maturation compartments, denoted as $Trasit$, and the circulation compartment $Circ$, which represent the observed circulating cells, see Figure 6.1.

The first differential equation in 5.2 describes the rate of the number of stem/progenitor cells change in the proliferation compartment. The rate of cell change is affected by the difference between the proliferation rate and elimination rate. If the generation rate is greater than the elimination rate, the proliferation rate will increase, otherwise reverse. Moreover, the proliferation and elimination rate can be influenced by the number cells in the proliferation compartment. More cells in bone marrow, the higher generation rate is. However, these stem and progenitor cells are sensitive to the drug. The drug is assumed to induce the cell division and generation and the magnitude of drug effect is highly dependent by the drug concentration, $E = slope \times Conc$ or E_{max} model, $Emax \times Conc/(EC50 * Conc)$. Additionally, a rebound cells from the circulation compartment can affect the proliferation rate called a feedback effect $(Circ_0/Circ)^{\gamma}$. The first differential equation is nonlinear with the rebound parameter, γ. The committed cell in the bone marrow goes through three transit compartments, and then can be observed in Circ compartment. If more cells go into a compartment and less go out, the rate of amount of cell will increase and otherwise decrease. The mechanism is the same for the last four compartments. Moreover, we also assume that the only cell loss is the cell "go" to next compartment, thus the proliferation rate, transit rate and the circulation rate should be equal at the steady state, i.e. $t = 0$.

A system of nonlinear differential equations are used to describe the whole mechanism process. The corresponding parameters include the baseline value of ANC ($Circ_0$), mean transit time (MTT), a drug related parameter (slope) and a feedback parameter (γ).

MTT represents the average time for a new generated cell to pass through the transit compartments before entering the circulation compartment. MTT has an inverse association with the transit rate, k_{tr}, defined as $MTT = (n+1)/k_{tr}$ where n is the number of transit compartments. Slope parameter is a fixed constant connecting drug concentration in the central compartment and drug effect. The parameter γ describes the strength of rebound cells from the circulation compartment.

$$\begin{aligned}
\frac{dProl}{dt} &= k_{prol}Prol(1-E_{drug})(\frac{Circ_0}{Circ})^{\gamma} - k_{tr}Prol \\
\frac{dTransit_1}{dt} &= k_{tr}Prol - k_{tr}Transit_1 \\
\frac{dTransit_2}{dt} &= k_{tr}Transit_1 - k_{tr}Transit_2 \\
\frac{dTransit_3}{dt} &= k_{tr}Transit_2 - k_{tr}Transit_3 \\
\frac{dCirc}{dt} &= k_{tr}Transit_3 - k_{circ}Circ
\end{aligned} \tag{5.2}$$

where $Prol, Transit_1, Transit_2, Transit_3$ and $Circ$ denote the amount of neutrophil count in the separate compartment. k_{prol}, k_{tr} and k_{circ} denote the proliferation rate, transit rate and circulation rate. The drug effect E is expressed by $Slope \times Conc$ in our application. At steady state, $\frac{dProl}{dt} = \frac{dCirc}{dt} = 0$, therefore, $k_{prol} = k_{tr} = k_{circ}$. Thus, the structure model parameters are $Circ0_i, MTT_i, \gamma_i, and Slope_i$ denoted by θ_i^{PD}.

We assume that the variability of observed ANC reflects the inter-subject variation and the measurement error expressed as,

$$y_{ij'}^{PD}|\theta_i^{PD} = f_i^{PD}(\theta_i^{PD}, \theta_i^{PK}, t_{ij'}) + \eta_{ij} \tag{5.3}$$

where f is a function of predicting ANC over time for subject i at time t_{ij}. The noise errors are assumed to follow the normal distribution independently respectively.

$$\epsilon_{ij} \sim N(0, \sigma_1^2) \qquad \eta_{ij'} \sim N(0, \sigma_2^2) \tag{5.4}$$

Given PK and PD model, the likelihood function is written as

$$p(y_i^{PD}, y_i^{PK}|\cdot) = \prod_{j=1}^{n_i} p(y_{ij}^{PD}|\theta_i^{PD}, \theta_i^{PK}, t_{ij'}, \sigma_2^2) \prod_{j=1}^{n_i} p(y_{ij}^{PK}|\theta_i^{PK}, t_{ij}, \sigma_1^2) \tag{5.5}$$

When we explored the data, we noticed that the shape of PD profiles are different and some shared the similar shapes in the population. The different shapes of PD curves may suggest investigate the heterogeneity of the population. Therefore, we proposed a nonparametric Bayes prior which induces a clustering property. See the clustering property in chapter 4.1. The goal of using a DP is to link the individualized PK and PD model and cluster the patients into groups. It helps us understand how drug works in the body.

$$\begin{aligned}(\theta_i^{PD}, \theta_i^{PK})|G &\sim G \\ G &\sim DP(\alpha, H) \\ H &\equiv N(\mu, \Sigma)\end{aligned} \tag{5.6}$$

where α is the concentration parameter and H is the base distribution of DP. The final stage includes additional hyper-parameters for σ^2, μ and Σ, respectively.

$$\sigma^2 \sim IG(a_0, b_0); \quad \mu \sim N(\mu_0, \Sigma_0); \quad \Sigma \sim IW(v_0, \Sigma_{iw}) \tag{5.7}$$

where IG denotes an inverse-gamma with mean $b_0/(a_0-1)$, IW denotes an inverse-Wishart distribution with mean $\Sigma_{iw}/(v_0 - p - 1)$, where p is number of parameters. The values of $a_0, b_0, v_0, \Sigma_{iw}, \mu_0$ and Σ_0 are fixed constants.

5.2 Euler Approximation

Using the differential equations is a common approach to describe a dynamic process in practice. It is also widely used in PK and PD models as it provides a time-varying rate of the response rather than the static average value. If the analytic solution of ODEs are

available, we can obtain the solution from the existing packages given a starting value of parameters. In practice, the analytical solution is often not available. For our case, it does not have an exact solution. Thus we use a numerical default function, ode in R. However, the simulation is pretty slow and the results are not stable either. We decided to use Euler's method to linearize the differential equations. Euler's approximation make the MCMC simulation 10 times faster than using the ODE solver in deSolve package. In addition, the results are very stable too. In the end, we implement Euler's method in the PD model. Here we show more details of the implementation.

In PD model 5.3, the values of $f_i^{PD}(\cdot, t_{ij'})$ require the solution of a system of ODEs given by 5.2. To keep the notations simple, we use the following equation to redefine the differential equations.

$$\frac{d\boldsymbol{v}}{dt} = g(t, \boldsymbol{v}(t, \boldsymbol{\theta}_i)) \quad \text{for} \quad t \neq t_0 \tag{5.8}$$

with $\boldsymbol{v}(t_0, \boldsymbol{\theta}) = \boldsymbol{v}_0(\boldsymbol{\theta})$. The vector $\boldsymbol{v}(\cdot) = (v_1(\cdot), \cdots, v_q(\cdot))^T$ represents the dynamic equations of q items, i.e. $Prol, Transit1, Transit2, Transit3, Circ$ in PD ODEs. $\boldsymbol{v}_0(\theta)$ is an initial condition. $\boldsymbol{g}(\cdot) = (g_1(\cdot), \cdots, g_q(\cdot))^T$ is the known function with respect to the parameters. The mean function $f_i^{PD}(\cdot)$ is directly related to $\boldsymbol{v}$, i.e. $f(\cdot) = H(\boldsymbol{v}(\cdot), t)$ where H is a known function.

We discretize the time points by an amount h, "step size", which is the distance between two consecutive time points. The observed time points may be unevenly distributed. We consider a discretization by N fixed time points $t_0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$ such that $t_{k+1} - t_k = h$ for $k = 1, 2, \cdots, (N-1)$. We choose the maximal value for these time points to be larger than t_n so that we can obtain the approximately fitted response at t_n. The solution to the ODEs can be expressed as

$$\boldsymbol{v}(t, \boldsymbol{\theta}) = \int_{t_0}^{t} g(s, \boldsymbol{v}(s, \boldsymbol{\theta}))ds + \boldsymbol{v}_0(\boldsymbol{\theta}) \tag{5.9}$$

which approximately equals to

$$\boldsymbol{v}(t+h,\boldsymbol{\theta}) - \boldsymbol{v}(t,\boldsymbol{\theta}) = \int_t^{t+h} g(s,\boldsymbol{v}(s,\boldsymbol{\theta}))ds \approx hg(t,\boldsymbol{v}(t,\boldsymbol{\theta})) \tag{5.10}$$

as $h \to 0$. Let $\tilde{\boldsymbol{v}_k} = \tilde{\boldsymbol{v}}(t_k,\boldsymbol{\theta})$ and $\tilde{\mu_k} = \tilde{\mu}(t_k,\boldsymbol{\theta}) = H(\tilde{\boldsymbol{v}_k})$ for $k = 1, 2, \cdots,$. We can write

$$\begin{aligned}\tilde{\boldsymbol{v}}(t_{k+1},\boldsymbol{\theta}) &= \tilde{\boldsymbol{v}}(t_k,\boldsymbol{\theta}) + (t_{k+1}-t_k)g(t_k,\boldsymbol{v}(t_k,\boldsymbol{\theta})) \\ \tilde{\mu}_k &= H(\tilde{\boldsymbol{v}}_k)\end{aligned} \tag{5.11}$$

with initial condition $\boldsymbol{v}_0(\boldsymbol{\theta})$. If $t_k \le t < t_{k+1}$, the value of approximate mean term at t is given

$$\tilde{\mu}(t,\boldsymbol{\theta}) = \tilde{\mu}(t_k,\boldsymbol{\theta}) + \frac{t-t_k}{t_{k+1}-t_k}(\tilde{\mu}(t_{k+1},\boldsymbol{\theta}) - \tilde{\mu}(t_k,\boldsymbol{\theta})) \tag{5.12}$$

Thus $f_i^{PD}(\cdot, t_{ij})$ is linearized by $\tilde{\mu}(t,\boldsymbol{\theta})$ function. This method is known as "naive" Euler's method, $\tilde{\mu_h}(t,\boldsymbol{\theta}) = \tilde{\mu}(t,\boldsymbol{\theta}) + o(h)$. In our simulation study, we compare the Euler's approach with a regular integrator. The results show that Euler's approximation improve the simulation speed. However, we need to determine the optimal size of step size, h. In addition, the optimal value of h is different across different patients. If h is too small, the approximation process stops at a certain time point before reaching t_N. In this case, we can not find the correspondingly fitted response. It is not possible to compute the likelihood function. If h is too large, we may have negative values which are meaningless in practice. Therefore, we need to choose h carefully. See more alternative linearization approaches such as "improved" Euler's method or Runge-Kutta method in Ghosh et al. (2011).

We also addressed another issue in our MCMC simulation. In high-dimensional case, it is difficult to optimize the tuning parameter to achieve efficient mixing. On one hand, we hope the chain is able to explore more regions. On the other hand, we also hope the proposed values are solvable in differential equations and are meaningful in practice. Therefore we implement the adaptive Metropolis which can automatically "learn" better parameter values of while the chain proceeds. We perform adaptive MCMC in both PK and

PD model. The proposal distribution is given at iteration n by $Q_n(\theta, \cdot) = N(\theta, 0.1^2 I_d/d$ for $n \leq 2d$; while for $n > 2d$,

$$Q_n(\theta, \cdot) = (1-\beta)N(\theta, \frac{2.38^2}{d}\Sigma_n) + \beta N(\theta, \Sigma_0) \tag{5.13}$$

where Σ_n is the covariance of empirical estimates $\theta_1, \cdots, \theta_n$, d is the dimension of parameter space, β is a small number between 0 and 1. Σ_0 is a fixed non-singular matrix which keeps the covariance matrix from collapsing to 0, i.e. $\Sigma_0 = 0.1^2 I_d/d$. The proposal distribution, $\mathrm{N}(\theta, 2.38^2\Sigma_n/d)$, is optimal in a high-dimension situation, Robert et al. (1997) and Roberts and Rosenthal (2001). In our work, we use a truncated sampling because some candidates, such as negative or extreme values, are ineligible in PK-PD modeling setting.

5.3 Predictive inference

There are several beneficial properties using the nonparametric bayesian model. We are able to tie the individual PK and PD model together and cluster the patients into several groups. We can also make full use of the full time course of concentration information. After performing the MCMC simulation, we are able to make inference of the clustering distribution and estimate the parameters by borrowing the strength from other patients. Additionally, we can do the prediction inference which is our greatest interest. We can predict the new patients' PD profile on the basis of their PK profile which is easier to obtain in practice. We use Polya Urn scheme, see more details in chapter 4, Dirichlet process, to relate the new observation to the $\theta_1^\star, \theta_2^\star, \cdots$ at each iteration. Then the predictive distribution of y_{n+1} is given by

$$\begin{aligned} p(y_{n+1}^{PD}|y_{1:n}^{PD}, y_{1:n}^{PK}) = \int\int & p(y_{n+1}^{PD}|\theta_{n+1}^{PD}, \theta_{n+1}^{PK}) \\ & p(\theta_{n+1}^{PD}, \theta_{n+1}^{PK}|\theta_{1:n}^{PD}, \theta_{1:n}^{PK}, y_{n+1}^{PK}, y_{1:n}^{PD}, y_{1:n}^{PK}) \\ & p(\theta_{1:n}^{PD}, \theta_{1:n}^{PK}|y_{1:n}^{PD}, y_{1:n}^{PK}) d\theta_{1:n}^{PD} d\theta_{1:n}^{PK} \end{aligned} \tag{5.14}$$

The middle term in equation 5.14 can be extended to

$$p(\theta_{n+1}^{PD}, \theta_{n+1}^{PK}|\cdot) \sim \frac{n_j}{n+\alpha}\sum_{j=1}^{k} p(y_{n+1}^{PK}|\theta_j^{PK}) + \frac{\alpha}{n+\alpha}p(y_j^{PK}|\theta_H^{PK}) \tag{5.15}$$

where k is the unique number of components at each iteration, n_j denotes the size of j^{th} cluster, θ_H^{PK} is a random draw from the base measure H. See more results in the simulation and real data studies.

5.4 Implementation of the MCMC Scheme to the PK/PD model

In order to implement the sampling scheme clearly, we start out by simplifying the notations. Let $\theta^\star = c(\theta_1^\star, \cdots, \theta_k^\star)$ denote the set of distinct θ_is and k, $(k \leq n)$ is the number of distinct elements. Let $(c_1, \cdots, c_n)$ denote the vector of configuration indicators, and n_j be the number of $c_i = j$. We implement Gibbs sampling by iterative sampling c_i from the full conditionals. The subscript "$-i$" indicates without the ith element of the vector. The superscript "$-$" refers to a summary with the appropriate observation or parameter removed. For example, k^- refers to the number of clusters formed by θ_{-i}. n_j^- represents the number of elements in cluster j when observation i is removed. All these notations work for both PK and PD.

The whole estimation sampling of mixture of DP model includes 4 steps. The first step is the initialization. The second step is implement Gibbs sampling with auxiliary components to sample the class configuration c_i for $i = 1, \ldots, n$. In third step, we use an adaptive Metropolis algorithm to estimate $\theta_j^\star$ for cluster j. The last step is to specify the hyper-parameters.

Step 1: The initial state

The permanent chain consists of the class configuration c_i, θ_i^{pk} and θ_i^{pd}. The initial values for the class label are 1 to n. Solving differential equations requires a "good" starting value to produce the reliable results. Thus, we set the initial values for θ_i^{PK} and θ_i^{PD} are the maximum likelihood estimates for PK and PD respectively. The parameters in the mixture model are initialized as follows. The base measure of DP, H, is a multivariate normal with mean parameter, the mean of MLEs, and covariance matrix, the covariance of MLEs. The precision parameter α is 1. The number of temporary auxiliary components is 3. The hyper-parameters are set to: $\mu_{pk} = \mu_{pd} = 0$, $\Sigma_0^{PK} = \Sigma_0^{PD} = I_d$, and inverse-Wishart with mean $\frac{\Sigma_{iw}}{v_0-p-1}$, $v_0 = 6$ $(v_0 > p+1)$, $\Sigma_{iw} = I_d$. $a_0 = 3$ and $b = 0.02$ in the inverse gamma distribution.

Step 2: Sample $[c_i|c_{-i}, \theta_{1:k^-}, \theta_{k^-+1}, \cdots, \theta_{k^-+m}]$

Repeatedly sample c_i for $i = 1, \cdots, n$ from I to IV:

- I. If $c_i \neq c_j$ for all $j \neq i$, then $c_i = k^- + 1$ and $\theta_{k^-+1}^{pk} = \theta_i^{pk}$, $\theta_{k^-+1}^{pd} = \theta_i^{pd}$. At the same time, draw the other two auxiliary parameters, $\theta_{k^-+2}^{pk}$ and $\theta_{k^-+3}^{pk}$ independently from H^{pk}; $\theta_{k^-+2}^{pd}$ and $\theta_{k^-+3}^{pd}$ independently from H^{pd}.

- II. if $c_i = c_j$ for some $j \neq i$, then $c_i = c_j$, $\theta_i^{PK} = \theta_j^{PK}$ and $\theta_i^{PD} = \theta_j^{PD}$. The class configurations for all auxiliary components are equal to $k^- + 1$, $k^- + 2$, $k^- + 3$. The three auxiliary parameters are randomly drawn from H^{pk} and H^{pd}.

- III. The corresponding probabilities in I and II are expressed as,

$$p(c_i = c|c_{-i}, y_i^{PD}, y_i^{PK}, \theta_{1:k^-}^{PK}, \theta_{k^-+1:k^-+m}^{PK}, \theta_{1:k^-}^{PD}, \theta_{k^-+1:k^-+m}^{PD}) \propto$$

$$\begin{cases} n_{-i,c} p(y_i^{PD}|\theta_c^{PD}, \theta_c^{PK}, \sigma_2^2) p(y_i^{PK}|\theta_c^{PK}, \sigma_1^2) & \text{for} \quad 1 \leq c \leq k^-, \\ \frac{\alpha}{m} p(y_i^{PD}|\theta_{auxiliary}^{PD}, \theta_{auxiliary}^{PK}, \sigma_2^2) p(y_i^{PK}|\theta_{au}^{PK}, \sigma_1^2) & \text{for} \quad k^- < c \leq k^- + m \end{cases}$$

- IV. After renewing each c_i for $i = 1, \ldots, n$, discard θ^{pk} and θ^{pd} which are not associated with any observations, $y_1, \cdots, y_n$. Keep the permanent terms, c_i, θ_i^{pk} and θ_i^{pd} for $i = 1, \cdots, n$.

In order to keep the notation simple in the following steps, let $\theta_1^{pk,\star}, \cdots, \theta_k^{pk,\star}$ denote the distinct values of $\theta_1^{pk}, \cdots, \theta_n^{pk}$, similar for PD, and $\theta_{1:k}^{pd,\star}$. The unique values of the class configuration is expressed as $c_1^\star, \cdots, c_k^\star$.

Step 3: Sample $[\theta_{pk}^\star|y_i, c_i = k]$ **and** $[\theta_{pd}^\star|y_i, c_i = k]$

In our model, we consider that $\theta_1^{pk,\star}, \cdots, \theta_k^{pk,\star}$ are i.i.d from $N(\mu_{pk}, \Sigma_{pk})$. Where $\mu_{pk} \sim N(\mu_0^{pk} = 0, \Sigma_0^{pk} = I_d)$ and $\Sigma^{pk} \sim$ Inverse-Whishart(v_0, Σ_{iw}^{pk}). Since these are conjugate priors, we can compute the full conditional distribution of $\mu^{pk}|\theta_{1:k}^{pk,\star}, \Sigma^{pk}$ and $\Sigma^{pk}|\theta_{1:k}^{pk,\star}, \mu^{pk}$. We used Gibbs sampler to sample $(\mu_{pk}^{1:T}, \Sigma_{pk}^{1:T})$ T times and set $\mu_{post}^{pk} = (1/T)\sum_1^T \mu_i^{pk}$ and $\Sigma_{post}^{pk} = (1/T)\sum_{i=1}^T \Sigma_i^{pk}$. See more details in chapter 4 section 2. It is similar for PD part $(\mu_{post}^{pk}, \Sigma_{post}^{pk})$. We have K clusters after step 2, then we update θ_k^{pk} and θ_k^{pd} in the following steps.

Repeat I and II for $k = 1, \ldots, K$. We take k^{th} cluster as example. Sampling $\theta_k^{pk}|y_{i\in k}^{pk}$ is as follows,

- I. The posterior distribution (target) is proportional to

$$p(\theta^{pk}|y_i^{pk}, c_i = k) \propto \prod_{i\in k} p(y_i^{pk}|\theta_k^{pk}) \pi(\theta_k^{pk}) \tag{5.16}$$

  Since there is no close form of likelihood, we use Metropolis-Hastings algorithms with an optimal proposal and an adaptive Metropolis algorithm in MCMC simula-

tion. Propose a candidate, $\theta^{\tilde{pk}}$ from 5.13 where θ is the current value of θ_i^{pk} and Σ_n is the covariance of all the unique estimates in the previous states. The final acceptance rate is,

$$\alpha(\theta^{pk}, \theta^{\tilde{pk}}) = \min\left\{1, \quad \frac{\prod_{i\in k} p(y_i^{pk}|\tilde{\theta}^{pk})\pi(\tilde{\theta}^{pk}, \mu_{post}, \Sigma_{post})}{\prod_{i\in k} p(y_i^{pk}|\theta_k^{pk})\pi(\theta_k^{pk}, \mu_{post}, \Sigma_{post})}\right\}$$

If the proposed values is accepted, we set $\theta_k = \tilde{\theta}$; otherwise, keep the current one.

- II. The steps of sampling $\theta_k^{pk}|y_{i\in k}^{pd}, y_{i\in k}^{pk}, \theta_k^{pk}$ are the same as $(3I)$ except the target density. It becomes,

$$p(\theta_k^{pd}|y_{i\in k}^{pd}, y_{i\in k}^{pk}, \theta_k^{pk}) \propto \prod_{i\in k} p(y_i^{pd}|\theta_k^{pd}, \theta_k^{pk}) \prod_{i\in k} p(y_i^{pk}|\theta_k^{pd})\pi(\theta_k^{pk}, \mu_{post}^{PD}, \Sigma_{post}^{PD})$$

Step 4. Re-sampling the hyper-parameters in the model

The typical mixture of Dirichlet process applications would include the hyper-priors. In each iteration, we do update the prior information given the associated observations are observed. For example, we update μ^{pk} and Σ^{pk} after we have the random sample $\theta_1^{pk}, \cdots,$ θ_k^{pk}. It is same as for μ^{pd} and Σ^{pd} too. In addition, we also update σ_1^2 and σ_2^2 given the Markov chain estimates in each iteration. We assume that the pk response $y_{ij}^{pk} \sim N(f_i^{pk}(\theta_i^{pk}, t_{ij}), \sigma_1^2)$ and $\sigma_1^2 \sim$ gamma$^{-1}(a_0, b_0)$. Given all the observations y_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, the posterior distribution of $\sigma_1^2|y_{ij}$ can be expressed as,

$$\begin{aligned} p(v|y^{pk}, \theta^{pk}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n_i} p(y_{ij}^{pk}|v)\pi(v) \\ &\propto (\frac{1}{v})^{\frac{N}{2}+a_0-1} \exp^{-\frac{1/2\sum_i\sum_j\left(y_{ij}-f_i^{pk}(\theta_i^{pk},t_{ij})\right)^2+b_0}{v}} \end{aligned}$$

where $v \equiv \sigma^2$. Therefore $\sigma^2|y \sim$ gamma$^{-1}(\frac{N}{2}+a_0, \frac{1}{2}\sum\sum(y_{ij}-f_{ij}^{pk}))^2+b_0)$. Similarly, we can sampling PD parameters from the posterior distributions. We update σ_1^2 and σ_2^2 at each iteration.

Repeat step 2, 3 and 4 to obtain a desired number of MCMC runs. There are more inference results in the simulation as well as real data.

5.5 A simulation study

In this section, we evaluated our proposed approach in a simulation. The simulation data is based on the real clinical trial study. We generated both PK and PD data for 30 patients. The "true" values of parameters are the maximum likelihood estimates of 3 patients who have different shapes of PK and PD profiles in the real data Figure 6.2. The noise were added, i.e. $\mathrm{N}(0, 0.01)$ for PK and $\mathrm{N}(0, 1)$ for PD, respectively. The variance is adopted to prevent the responses to be negative or too close to 0. We kept the well-balanced and non-negative response values.

The system of nonlinear differential equations do not have an exact solution. We used deSolve, a R package, to solve the ODEs at first. However, it is slow and the results are unstable due to the complex structure of ODE system. A MCMC simulation with 1000 iterations were tested using one of most popular numerical solver of ODEs in deSolve package. On average, it takes more than 6 minutes to complete an iteration. To compare, Euler's method was tested under the same setting. It took 0.63 minute for one iteration. Thus Euler's approach is almost 10 times faster than lsoda solver. Figure 6.3 and 6.4 show that the mean of PD parameter estimates with 95% probability interval. Each column represents the parameter mean estimates for one patient. The estimates for both approaches are very close to each other for each patient. As ODEs were needed to solve more than 1000 times in each iteration, Euler's method was adopted for the following MCMC simulation. Moreover, we implement a combination of Metropolis-Hastings algorithm and a Gibbs sampling with auxiliary components in the MCMC simulation. See more details about MCMC algorithm in chapter 5, section 4.

Figure 6.5 shows the distribution of the clustering distribution of the maximum likelihood estimates. Our approach indicates that there are 3 groups among 30 patients because it has a relative large weight. This result is consistent to the true setting. Figure 6.4 predict the individual parameter estimate with 95% probability interval. We can see that the first ten patients' parameter estimation as well as the confidence interval are very close to each other. It suggests the first 10 patients belong to a cluster. Similarly, the last 10 patients should go into second group. The rest belongs to the third cluster. At the same time, PK parameter estimates show the similar pattern too, Figure 6.6. The Dirichlet Process is best used for density estimation, and, as it is our case, for borrowing strength across observations and improving estimation when the curves need to be estimated with sparse data. This appears to be overall achieved by our model.

Our approach with Euler approximation can improve the estimates across the cycles. Figure 6.7 shows three PD curves fitted with an ODE solver package in R, and the "true curve" that are used to generate the data as well as the final simulated data points (cross signs). Figure 6.8 present the PD curves with Euler's estimates for three patients in the simulation data. The fitted values are estimated by $f_i^{PD}(t) = (\hat{\theta}_i^{pd}, t)$, where $\hat{\theta}_i$ is the posterior mean from MCMC samples.

Figure 6.9 shows that the PD response for a new patient are estimated by $\hat{f}^{PD}(t) = 1/T \sum f^{PD}(\theta_{n+1}^{pd}, t)$, where θ_i^t is the t-th imputed parameter vector θ_i in simulated Monte Carlo posterior sample. The interval is 95% probability interval of the predicted response for each observed time point. In each iteration, we use $p(y_{n+1}^{PK}|\theta_{1:n}^{PK}, \theta_{auxiliary}^{PK}, \sigma_1)$ to determine the new patient going into which clusters. If the new patient goes into "true" cluster which are used to generate the new patient data, the curve will be completely consistent with the data. However, it is possible that the new patient will be assigned in one of two existing clusters or the new cluster. In this case, the predicted curve will be not consistent with the data as we expected. The final predictive distribution might heavily depend

on the cluster assignment in the estimation step. The mean response at each observed data point is an average value of all predictions from different cluster across iterations. For a new patient generated from ID 57 occasion 1, there is 86.4% iteration that our model assigns the new patient to the "true" curve, 1.4% iteration that the new patient goes into the new cluster, 7.2% iteration that it is assigned to the first cluster which generating the first ten cycles, 28.4% iteration that it goes to the second cluster which simulates the middle ten cycles. See results in Figure 6.9.

In addition, we check the trace plot of a patient's PD parameters, Figure 6.10. The plat line ("gap") in the plot indicates that the proposed candidates is meaningless so we keep the current state values.

5.6 Real data example

Here we use a real clinical trail data, Friberg (2002), to assess the performance of our proposed approach. Usually, the clinical data has a precise administration. Overall 45 patients with different cancer forms only received paclitaxel, a single anticancer agent in 196 cycles (varying between one and 18 cycles per patient, median, 3 cycles). Paclitaxel was administered as a 3-hour infusion, with an initial dose of 175 mg/m^2 every 3rd week. Unbound plasma concentrations were monitored on course 1 and course 3, with an average of 3.5 samples per patient and course.

We chose 40 cycles from 45 patients. These 40 cycles satisfy two conditions. First, the cycle has PK and PD data. Second, PD cycle has three or more observations. Figure 6.11 shows two fitted curves in the real data. The parameter initialization is the same as those in the simulation study. The following estimation and prediction results are based on 5000 MCMC simulation.

Figure 6.12 shows that the PD response for a new patient are estimated by $\hat{f}^{PD}(t) = 1/T \sum_{t=1}^{T} f^{PD}(\theta^{pd,t})$, where $\theta^{pd,t}$ is the t-th imputed parameter vector in simulated Monte Carlo posterior sample. The interval is 95% probability interval of the predicted response for the observed time point.

Chapter 6

Discussion

We provide a coherent probability model for the analysis of PK/PD mechanistic model. By using a bayes approach, we are able to tie individualized PK and PD cycles together, obtain the inference on patients' clustering on the basis of their concentration or response profiles. The cluster configuration allows us to predict the new patients' drug response based on its drug concentration data set. This approach helps us to investigate the heterogeneity of the population and then provides us a chance to gain more information from each subpopulation. This model not only can combine prior information with the clinical trial data, but also deal with complex dynamic systems. Thus the results of estimated dynamic parameter based on this model should be reliable and reasonable to interpret long-term PK/PD dynamics. This approach can help us better understand how the drugs work in the body first in the subpopulation then in the whole population.

Although nonparametric Bayesian model is a promising perspective to address the population PK/PD models, there are several challenges too. First, our approach is highly computationally intensive. Our model requires solving a system of nonlinear differential equations. It becomes very difficult when we have a limited data in a few cycles. That is also one of the reasons that we introduced the Euler's method to approximately estimate

the corresponding response. Another challenge is that how to set up a reasonable truncated the proposal distribution so that we can allows the chain to explore a large region and also propose the doable parameters which is solvable in the highly unstable ODEs system. More importantly, the proposed values are physiologically meaningful.

There are two directions I am interested in the future. One is to improve our approach by incorporating the available covariates information including body surface area, bilirubin (hematoidin, excreted in bile), genetic biomarkers as well as the basic demographic variables. The second direction is about missing data. We want to provide a reasonable approach to deal with the "missing" data since it is a fairly common issue in clinical trial. In real example, we only have one or two data points for some PD cycles. We are not very clear how to do the parameter estimation. Right now, many software packages either throw away all subjects who have missing data or impute missing values with population mean or some fixed values. The first approach throws away a potential large amount information, where as Hoff et al.(2009) proposed to use Gibbs sampling for the missing data. Tsiatis et al.(2006) discussed the nonparametric theory and missing data which provides us the framework to do deep research. In addition, we plan to write a R package for our approach.

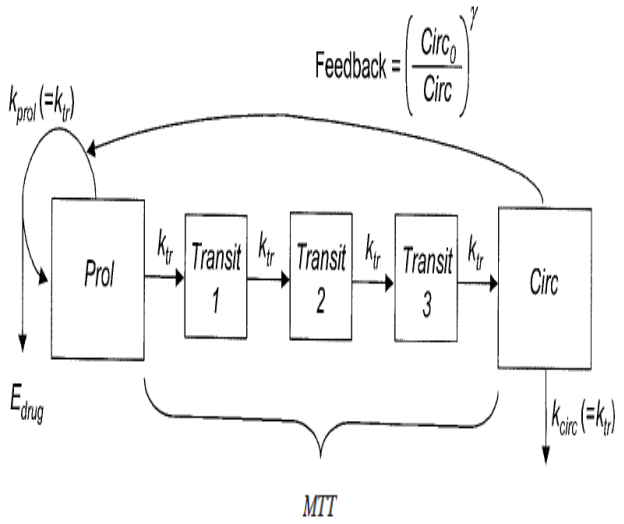

Figure 6.1: Neutrophil cell proliferation model with feedback. Stem or progenitor cell compartment, Prol; Maturation compartment for transit including Transit 1, Transit 2, Transit 3. circulating neutrophil compartment, Circ. E_{drug} represents drug effect. Feedback represents the strength of rebound cells. MTT measures the mean transit time. k_{prol}, k_{tr} and k_{circ} represent the proliferation rate, transit rate and circulation rate, respectively.

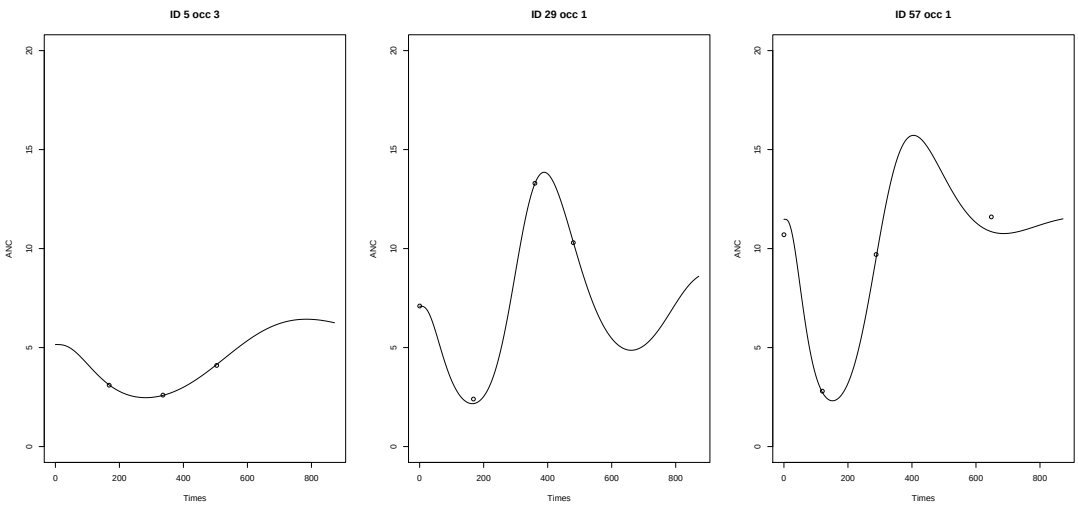

Figure 6.2: The fitted curves chosen for generating the simulation data.

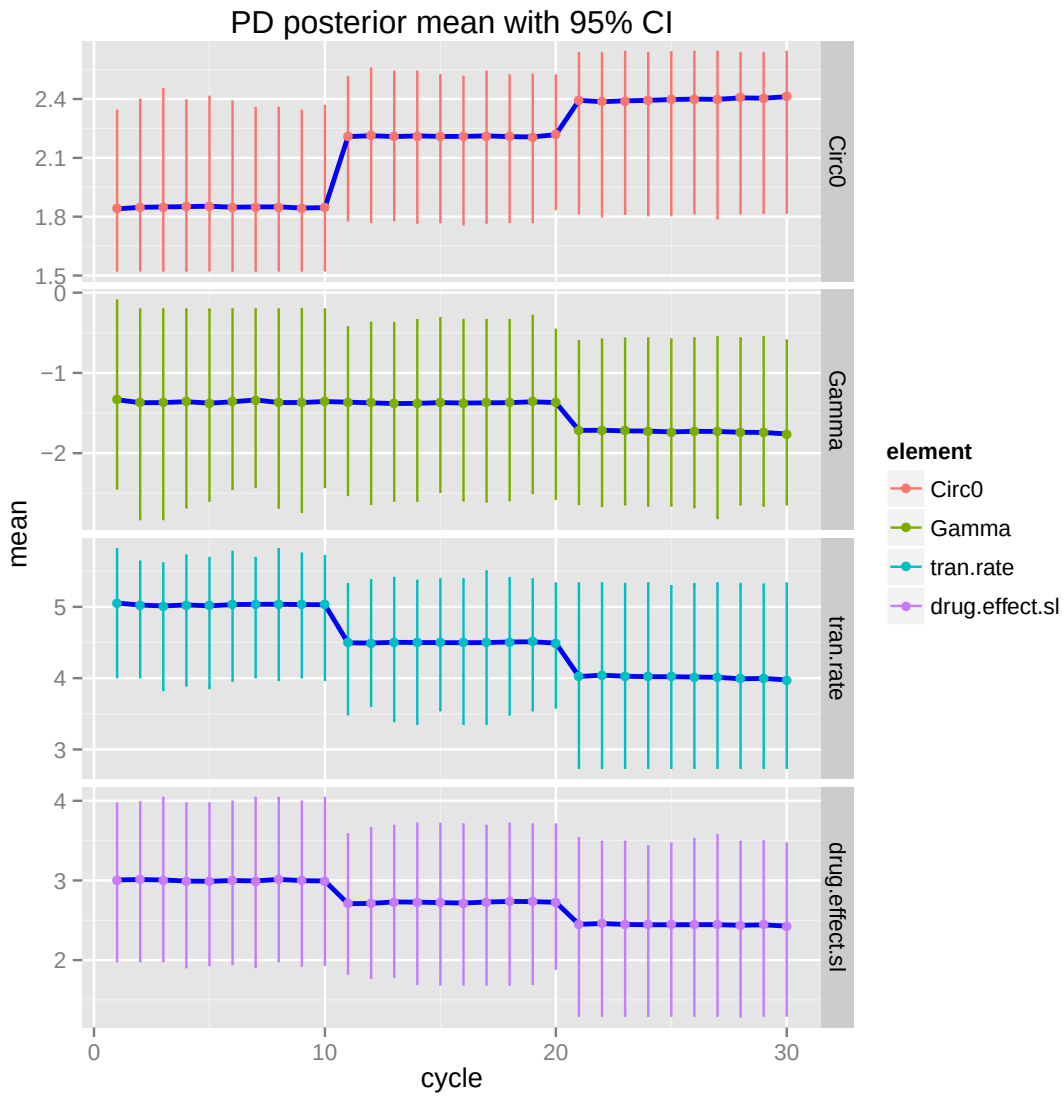

Figure 6.3: PD parameter estimates: posterior mean with 95% probability interval across patients using Euler linearization.

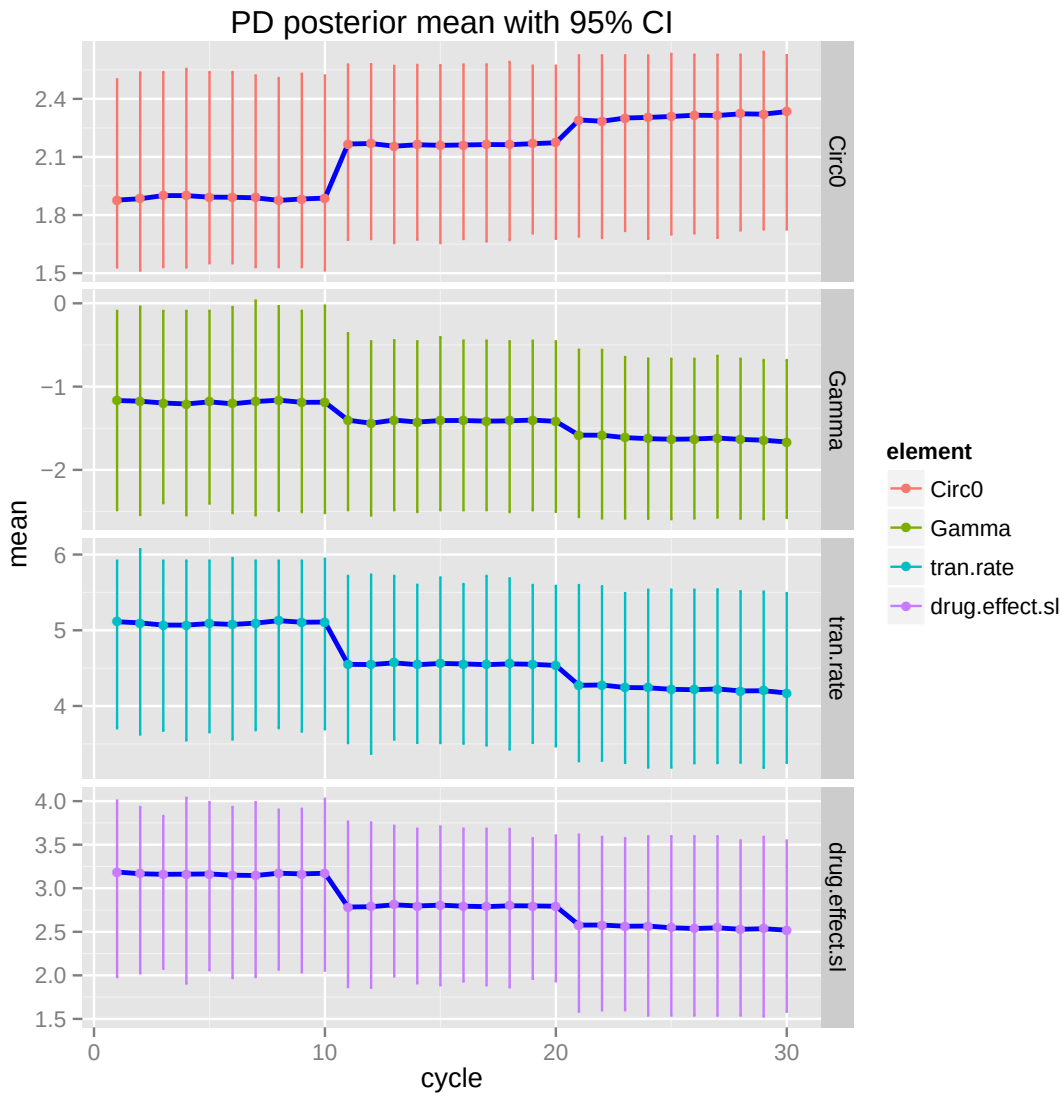

Figure 6.4: PD parameter estimates: posterior mean with 95% probability interval across patients using lsoda solver.

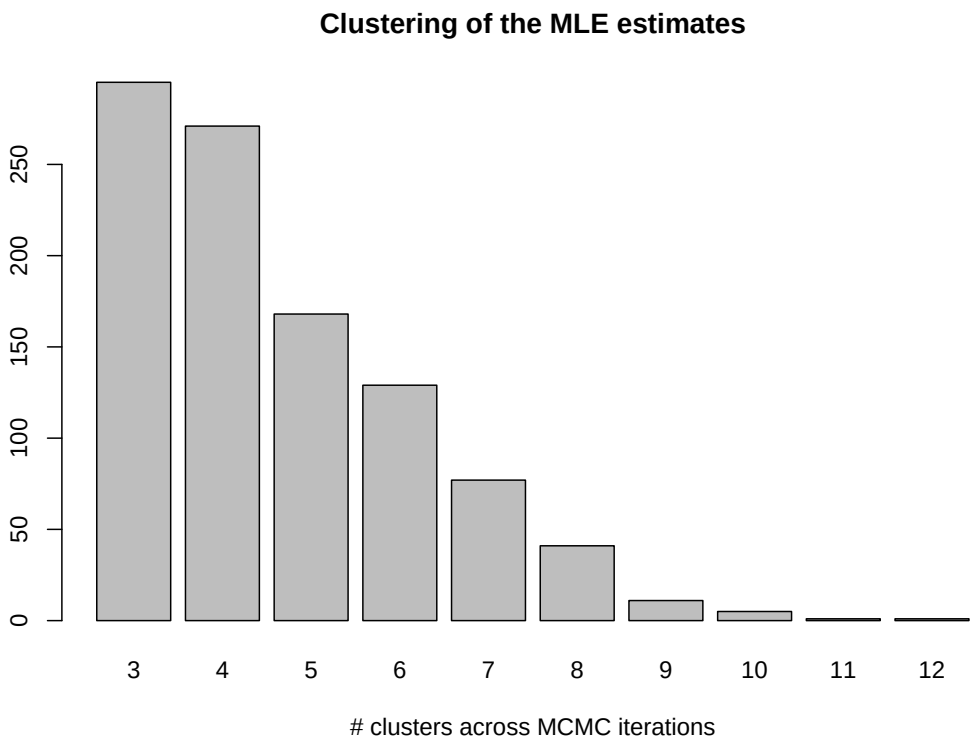

Figure 6.5: Clustering inference from Euler's method for the simulation data.

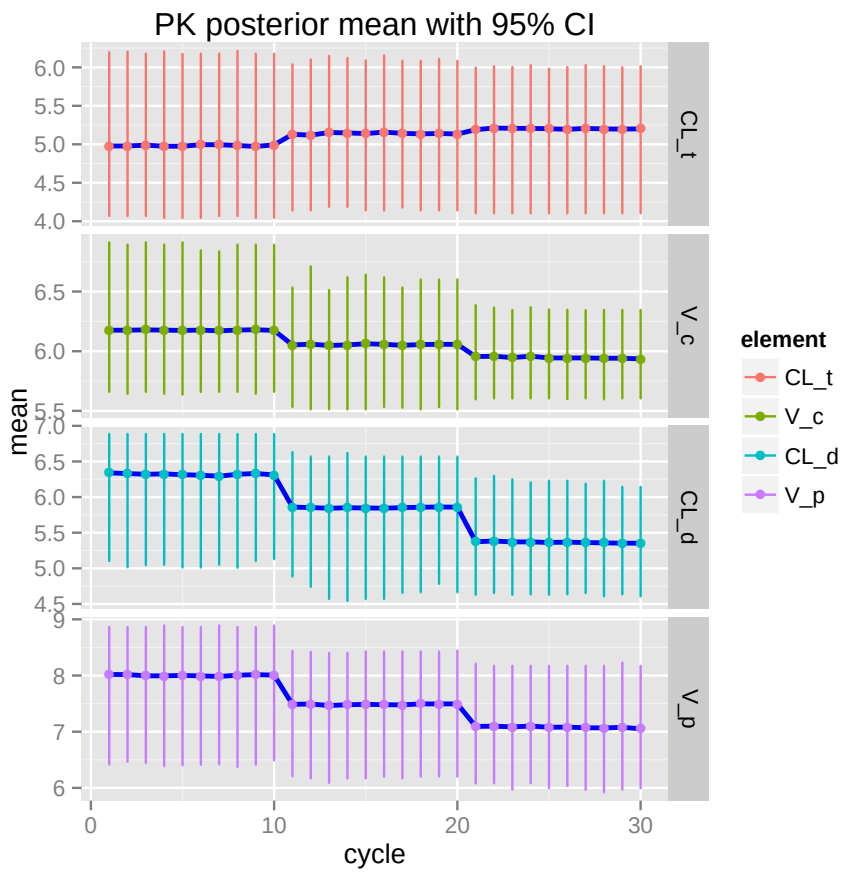

Figure 6.6: PK parameter estimates: posterior mean with 95% probability interval across patients.

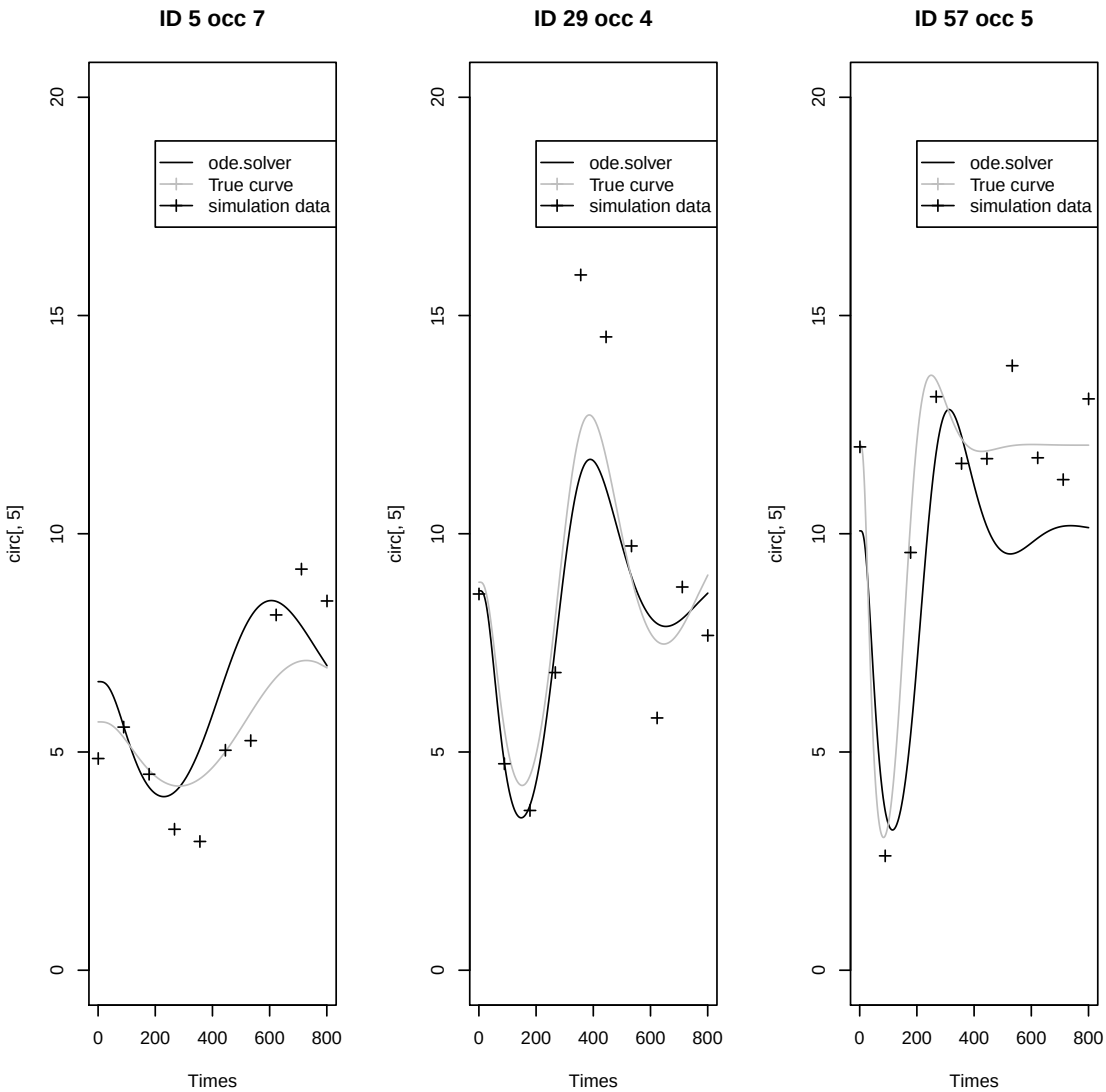

Figure 6.7: Fitted curves with MLEs for three patients in the simulation study.

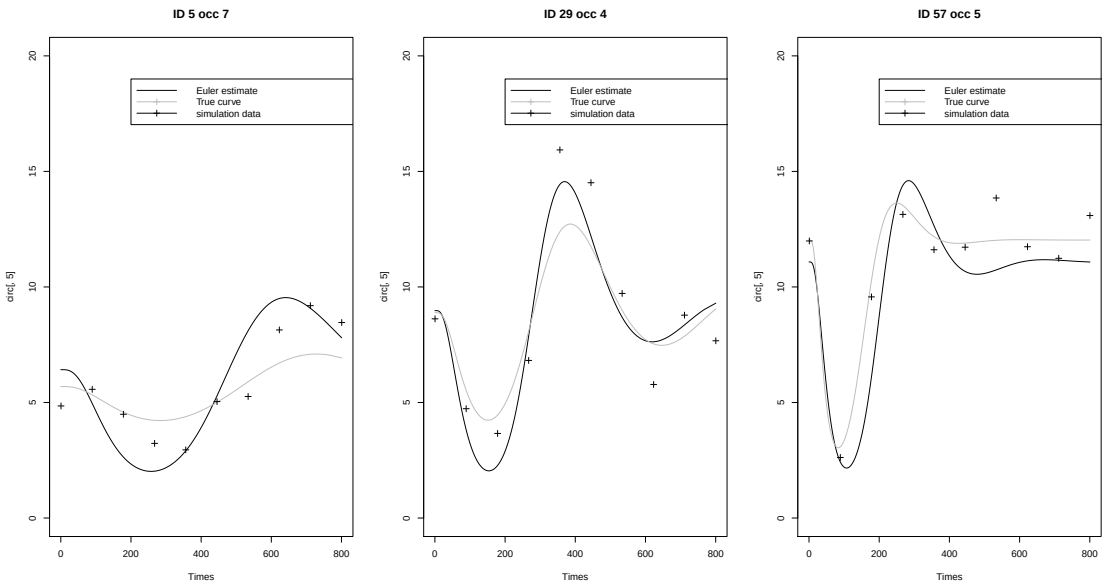

Figure 6.8: Fitted curves with posterior mean for three patients in the simulation study.

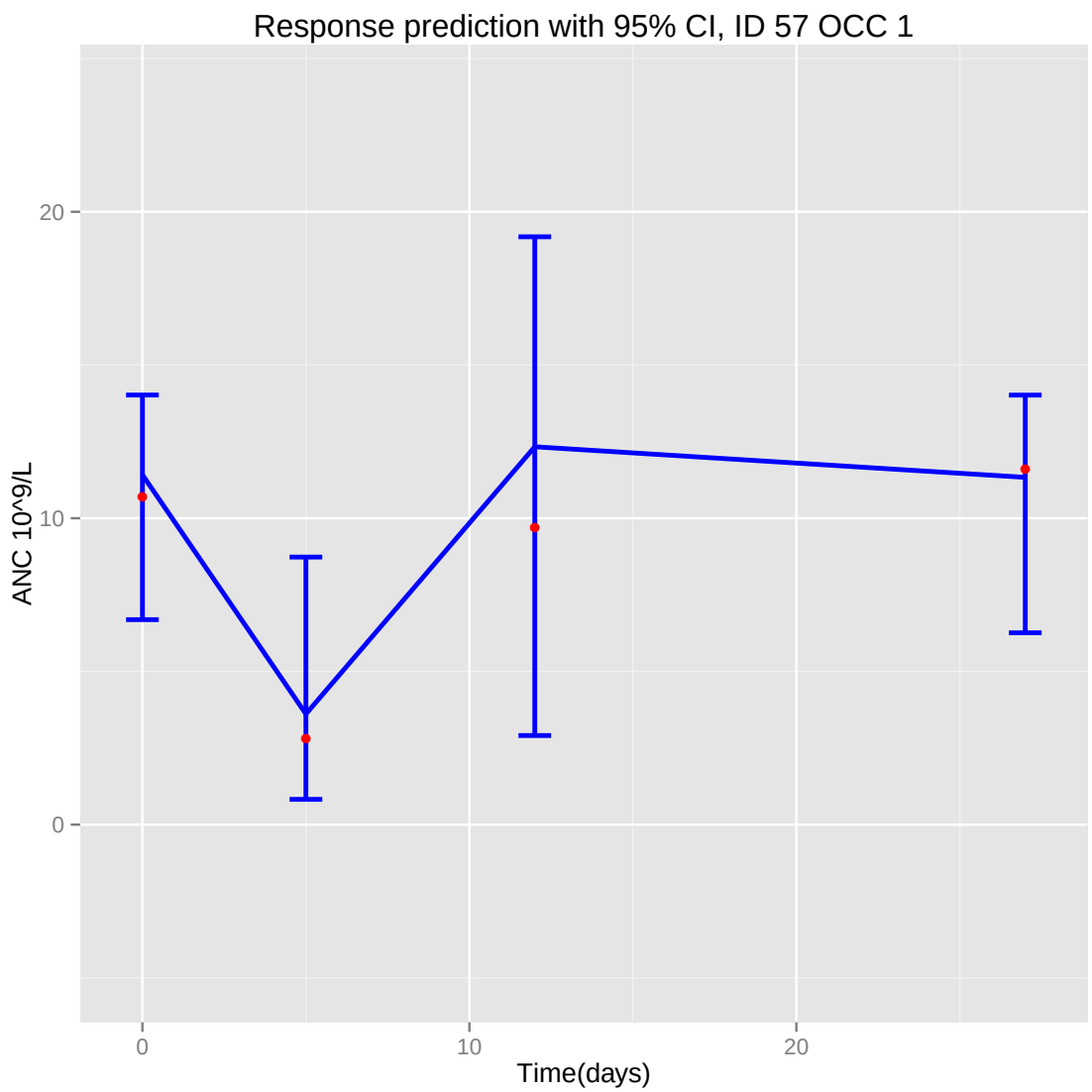

Figure 6.9: Predicted absolute neutrophil count for a new patient in the simulation study.

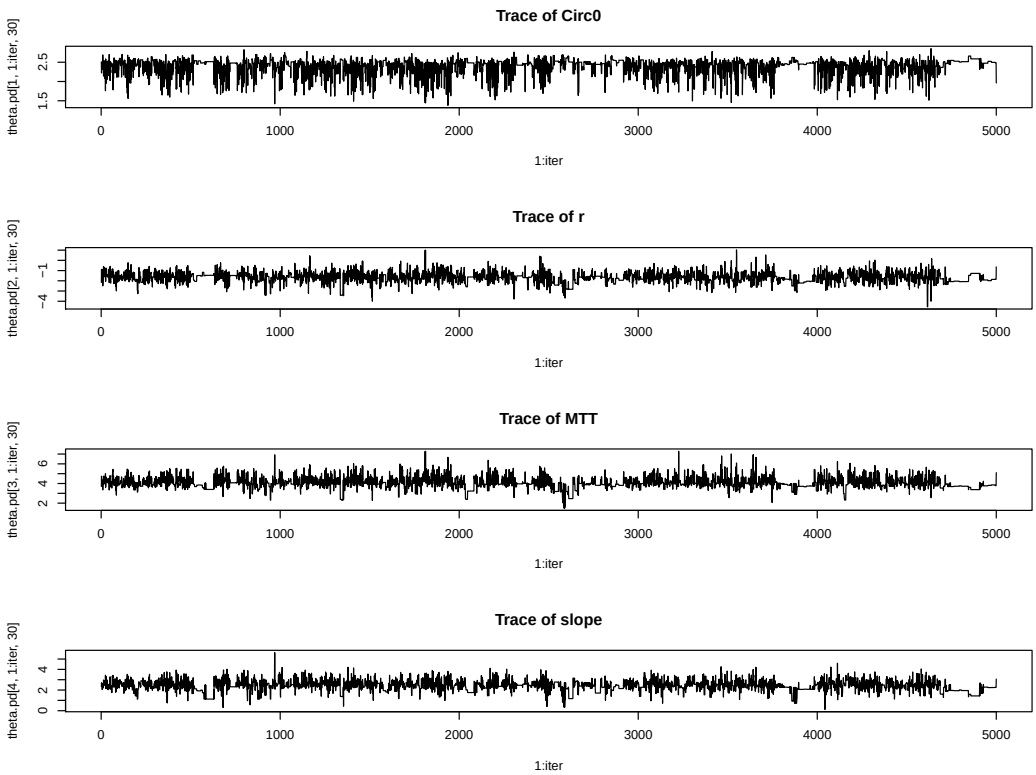

Figure 6.10: Trace plot of PD parameters across the iterations.

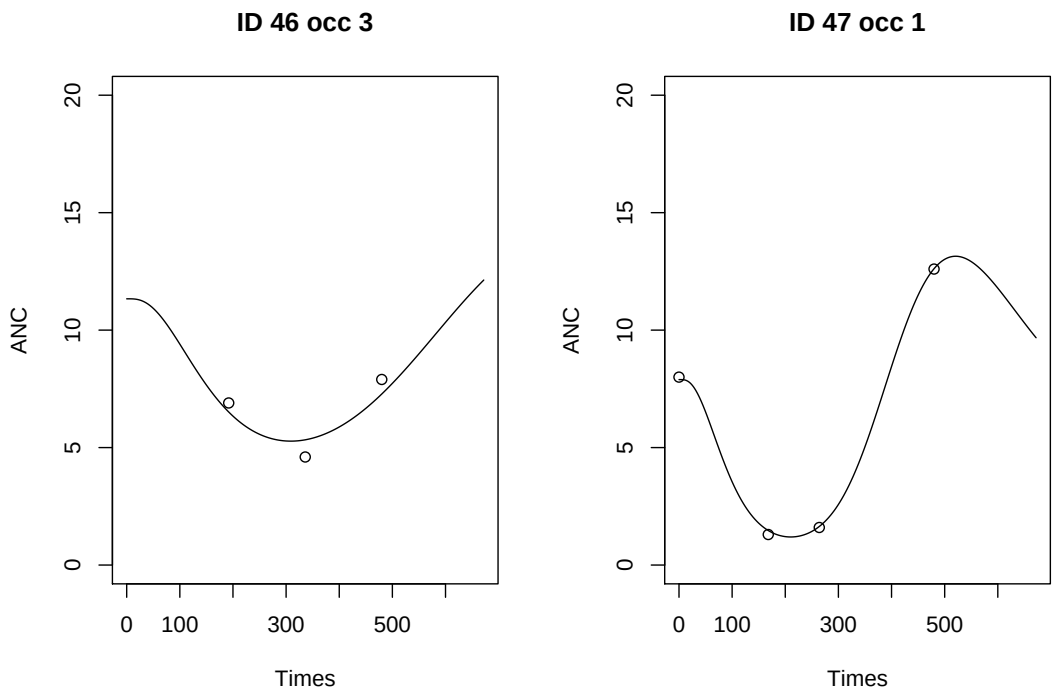

Figure 6.11: The fitted curves from the real data set.

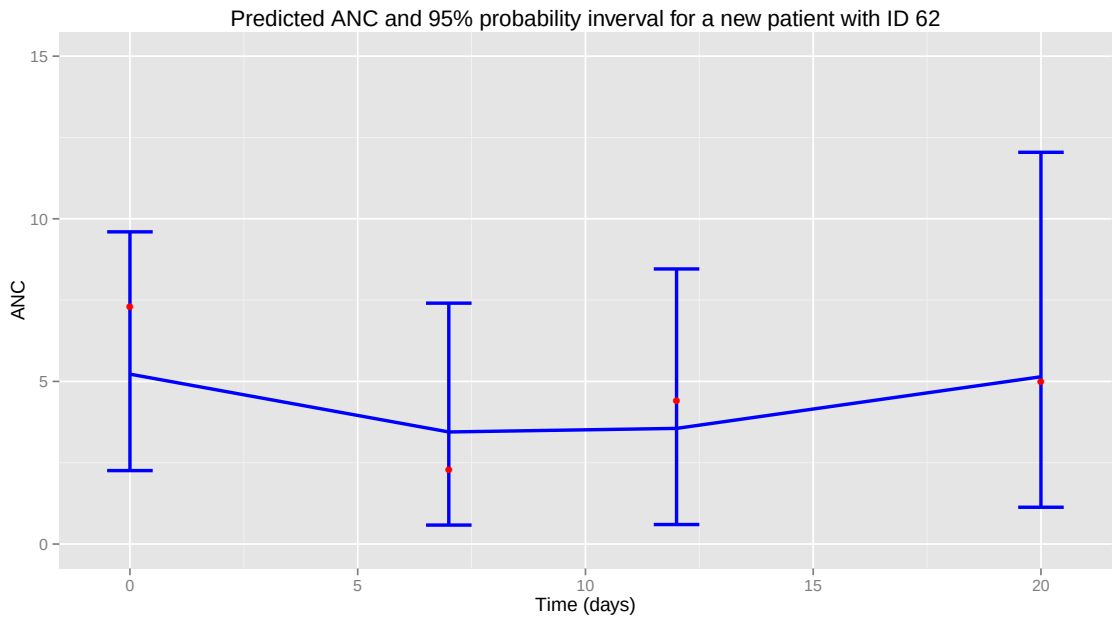

Figure 6.12: Predicted absolute neutrophil count for a new patient in the real data.

Bibliography

[1] Andrieu C. and Roberts C.P.(2001) Controlled MCMC for optimal sampling. Technical report, University Paris Dauphine.

[2] Andrieu C., Moulines E. and Priouret P.(2005) Stability of stochastic approximation under verifiable conditions. *Siam Journal on control and optimization* 44:283-312

[3] Andrieu C. and Atchade Y.F.(2007) On the efficiency of adaptive MCMC algorithms. *Electronic Communications In Probability*, 12:336-349

[4] Andersen M.E., Clewell H.J., Gargas M.L., Smith F.A., Reitz R.H. Physiologically based pharmacokinetics and the risk assessment process for methylene chloride. *Toxicol Application Pharmacology* 87: 185-205 .

[5] Aronson J.K. (2005). Moving about. *Br. J. Clin. Pharmacology* 59: 263-4.

[6] Bai Y. (2008) Simultaneous drift conditions on adaptive Markov Chain Monte Carlo algorithms. *Techinical Report in Department of Statistics at the University of Toronto*

[7] Bai Y., Roberts G.O. and Rosenthal J.S. (2009) On the contanment condition of adaptive Markov Chain Monte Carlo algorithm. *Techinical Report in Department of Statistics at the University of Toronto*

[8] Brooks S.P. and Gelman D.B (1998) Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7, 434-455.

[9] Benowitz N., Forsyth R.P., Melmon K.L., Rowland M.(1974) Lidocaine disposition kinetics in monkey and man. I. Prediction by a perfusion model. *Clin Pharmacol Ther* 16: 8798.

[10] Bischoff K.B., Dedrick R.L., Zaharko D.S., Longstreth J.A. (1971) Methotrexate pharmacokinetics. *J. Pharmacology Science* 60: 1128-33.

[11] Brightman F.A., Leahy D.E., Searle G.E., Thomas S. (2006). Application of a generic physiologically based pharmacokinetic model to the estimation of xenobiotic levels in rat plasma. *Drug Metab Disposition* 34: 84-93.

[12] Brightman F.A., Leahy D.E., Searle G.E., Thomas S. (2006). Application of a generic physiologically based pharmacokinetic model to the estimation of xenobiotic levels in human plasma. *Drug Metab Disposition* 34: 94-101.

[13] Bonate P.L.(2011), Advanced reveiw, Pharmacokinetics. *John Wiley and Sons, Inc* Vol. 3

[14] Bennett J.E., Racine P.A. and Wakefield J.C.(1996). Markov chain Monte Carlo for nonlinear hierarchical models, in Markov chain Monte Carlo in Practice. Chapman and Hall, London, 339357.

[15] Boxer G.E. and Jelinek V.C. (1948) Streptomycin in the blood: Chemical determinations after single and repeated intramuscular injections. *J. Pharmae. exp. Ther.* 92: 226-235.

[16] Beal S.L. and Sheiner L.B. (1982) Estimating Population Kinetics. *CRC Crit. Rev. Biomed. Eng.* 8, 195-222.

[17] Blackwell D. and MacQueen J.B.(1973). Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1:353355

[18] Carlin B.P. and Louis T.A. (2000). Bayes and Empirical Bayes Methods for Data Analysis. London: Chapman and Hall, 2nd edition.

[19] Craigmill A.L. (2003) A physiologically based pharmacokinetic model for oxytetracycline residues in sheep. *J. vet. Pharmacol. Therap.* 26, 5563.

[20] Chien J.Y. and Mohutsky M.A. (2003) Physiological approaches to the prediction of drug-drug interactions in study populations. *J. Curr Drug Metab* 4: 347-356.

[21] Cascone S., Lamberti G., Titomanlio G. and Piazza O. (2013), Pharmacokinetics of Remifentanil: a three-compartmental modeling approach. *Translational modeicine UniSa*. 7: 18-22

[22] Bauer R.J., Guzy S. and Ng C. (2007) A survey of population analysis methods and software for complext pharmacokinetic and pharmacodynamic models with examples. *J. Pharmacokinetic Pharmacodynamic* 33:227-279

[23] Calderhead B., Girolami M. and Lawrence L.D.(2009). Accelerating Bayesian inference over Nonlinear differential equations with Gaussian processes *Advances in neural information processing systems,* 21, 217-224

[24] Cowles M.K. and Carlin B.P. (1996) Marlov chain Monte Carlo convergence diagnostics: a comparative review, J. American Statistics Association*Scandinavian Actuarial Journal* 36, 24-63.

[25] Davis N.R. and Mapleson W.W. (1993) A physiological model for the distribution of injected agents, with special reference to pethidine. *Br J Anaesth* 70: 24858.

[26] Dong M., Fukuda T., Cox S., De V.M., Hooper D.K., Goebel J. and Vinks A.A. (2014) Population pharmacokinetic-pharmacodynamic modelling of mycophenolic acid in paediatric renal transplant recipients in the early post-transplant period. *Br J Clin Pharmacol* 78(5):1102-12.

[27] Duffull S.B., Waterhouse T.H. and Eccleston J.A. (2005) Some considerations on the design of population pharmacokinetic studies. *Journal of Pharmacokinetics and Pharmacodynamics* 32:441457.

[28] Derendorf H. and Meibohm B. (1999), Modeling of Pharmacokinetic/Pharmacodynamic relationships: Concepts and Perspectives. *Pharmaceutical Research* Vol.16, No.2

[29] Dickschen K., Willmann S. and etc (2012), Physiologically Based Pharmacokinetic Modeling of Tamoxifen and its Metabolites in Women of Different CYP2D6 Phenotypes Provides New Insight into the Tamoxifen Mass Balance. *Front Pharmacol* 3: 92.

[30] DiMasi J., Hansen R. and Grabowski (2003), The price of innovation: New estimates of drug development cost. *J. Health Economics* 22, 2003.

[31] DiMasi J., Hansen R. and Grabowski (2001), New Drug Development in the United States from $1963 - 1999$. *Clinical Pharmacology and Therapeutics* 69, No.5 : 286-296.

[32] Davidian M. and Giltinan D.M. (1995). Nonlinear Models for Repeated Measurement Data. London: Chapman and Hall.

[33] ESCOBAR M.D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* 89, 268277.

[34] ESCOBAR M.D. and WEST M. (1995). Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association* 90, 577588.

[35] ESCOBAR M.D. and WEST, M. (1998). Computing nonparametric hierarchical models. *In Practical nonparametric and semiparametric Bayesian statistics*, pp. 122.

[36] Friedman H., Greenblatt D.J., Peters G.R., Metzler C.M., Charlton M.D., Harmatz J.S., Antal E.J., Sanborn E.C. and Francom S.F. (1992) Pharmacokinetics and pharmacodynamics of oral diazepam: effect of dose, plasma concentration, and time. *Clinical Pharmacology Therapy* 52(2):139.

[37] Guidance for industry: Bio-availability and bio-equivalence studies for orally administered Drug Products General Considerations (2000) US department of health and human services, Food and Drug Administration.

[38] Ferguson T.S. (1973). A Bayesian Analysis of SomeNonparametric Problems. *The Annals of Statistics* 1, pp. 209-230.

[39] Gaudino M. (1949) Kinetics of distribution of inulin between two body water compartments. *Proc. Soc. exp. Biol. Med.* 70: 672-674.

[40] Gabrielsson J. and Weiner D.(2012), Noncompartmental Analysis. *Computational Toxicology, methods in molecular biology* Volume 929, pp 377-389

[41] Geman S. and Geman D. (1984) The price of innovation: New estimates of drug development costStochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern Recognition and machine Intelligence*6, 721741.

[42] Gelman A. and Rubin D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science,* 7, 457-511.

[43] Pasarica C. and Gelman A.(2010) Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*

[44] Gelman A., Bois F. and Jiang J. (1996) Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* 91, 14001412.

[45] Gillespie W.R. (1991) Noncompartmental versus compartmental modelling in clinical pharmacokinetics. *clin. Pharmacokinet* 20(4):253-62

[46] Grass G.M. and Sinko P.J. (2002) Physiologically-based pharmacokinetic simulation modelling. *Adv. Drug Deliv Rev* 54:433-451

[47] Ghosh K.S. and Goyal L. (2011) Statistical inference for Non-linear models involving ordinary differential equations. *Journal of statistical theory and practice* 4(4)

[48] Gilks W.R. and Wild P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 41, pp 337-348.

[49] Haario H., Saksman E., and Tamminen J. (2001) An adaptive Metropolis algorithm. *Bernoulli* 7, 223-242.

[50] Hastings W.K. (1970) Monte-Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57(1)

[51] Holte S.E., Cornelisse P. and Self S. (2003) An alternative to nonlinear least-squares regression for estimating parameters in ordinary differential equations models.

[52] Huang S.M. and Rowland M. (2012) The role of physiologically based pharmacokinetic modeling in regulatory review. *Clin Pharmacol Ther.* 91(3):542-9

[53] Johnson T.N. (2005) Modelling approaches to dose estimation in children. *Br. J. Chin Pharmacol.* 59:663-669

[54] Jbabdi S., Woolrich M.W. and Behrens TEJ. (2009) Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models *NeuroImage* 44, 2, 373384.

[55] Kreuer S., Hauschild A., Fink T., Baumbach J.I. and Maddula S. (2014) Two different approaches for pharmacokinetic modeling of exhaled drug concentrations. *Sci. Rep.* 4:5423.

[56] Kim S., Tadesse M.G. and Vannucci M. (2006) Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4), 877893.

[57] Jurgen B.B., Ayhan B., Beom S.S. and Cornelia B.L. (2011) Development of a New Pre- and Post-Processing Tool (SADAPT-TRAN) for Nonlinear Mixed-Effects Modeling in S-ADAPT. *AAPS* 13(2):201-211

[58] Lalonde R., Straka R., Pieper J., Bottorff M. and Mirvis D. (1987), Propranolol pharmacodynamic modelling using unbound and total concentrations in healthy volunteers *J Pharmacokinet Biopharma* l5: 569-582.

[59] Lunn D.J. and Aarons L.J. (1997), Markov chain Monte Carlo techniques for studying interoccasion and intersubject variability: Application to pharmacokinetic data. *J. of Applied Statistics*. 46(1): 73-91

[60] Lunn D.J. and Aarons L.J. (1998), The pharmacokinetics of saquinavir: A Markov chain Monte carlo population analysis. *J. of Pharmacokinetics and Biopharmaceutics*. 26(1): 47-74

[61] Lunn D.J., Wakefield J., Best T.A., Wakefield J. and Spiegelhalter D. (1999), PKBUGS User Guide, *J. Pharmacokinet. Biopharmaceut*. 3457-478

[62] Lunn D.J., Wakefield J., Best T.A., Wakefield J. and Spiegelhalter D.(2002), Bayesian analysis of population PK/PD models: General concepts and software. *Journal of Pharmacokinetics and Pharmacodynamics*. 29(3): 271-307

[63] Lunn D.J. (2004) WinBugs Differential Interface. Department Epidemiology and Public Health, Imperial College School of Medicine

[64] Lin J.(1995) Species similarities and differences in pharmacokinetics. *Drug Metab. Dispos.* 23:1008-1021

[65] Lupfert C. and Reichel A. (2005) Development and application of physiologically based pharmacokinetic-modeling tools to support drug discovery.*Chem. Biodversity* 2:1462-1486

[66] Lindstrom M.L. and Bates D.M. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics* 46(3):673-87.

[67] MacEachern S.N. (2000) Dependent Dirichlet Processes, Technical Report, Department of Statistics, The Ohio State University

[68] MacEachern S.N. and Muller P. (1998) Estimating mixture of DP models, *Journal of Computational and Graphical Statistics*. 7, 223-238.

[69] Mimmelstein K.J. and Lutz R.J. (1979). A review of the applications of physiologically based pharmacokinetic modeling. *J. Pharmacokinet Biopharm* 7: 127-45.

[70] Mengersen K.L, Robert C.P. and Guihenneuc-Jouyaux J. (1999) MCMC convergence diagnostics: a review *Bayesian Statistics* 6.

[71] Mager DE., Wyska E. and Jusko WJ. (2003) Diversity of mechanism-based pharmacodynamic models. *Drug Metab Dispos*. 31(5):510-8.

[72] Reddy M., Yang RS., Andersen ME. and Clewell HJ. (2005) Physiologically Based Pharmacokinetic Modeling: Science and Applications. Published by John Wiley and Sons, Inc.

[73] MULLER P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* 83, 6779.

[74] Ngashima R, OReilly RA, and Levy G. (1969), Kinetics of pharmacologic effects in man: the anticoagulant action of warfarin. (1969), *Clinical Pharmacology Therapy,* 10:22-35.

[75] Nielsen E. and Friberg L.(2013). Pharmacokinetic-Pharmacodynamic Modeling of Antibacterial. *Drugs.Pharmacological Reviews*, 65(3): 1053-1090

[76] Neal R. (1997), Markov Chain Monte Carlo Methods Based on Slicing the Density Function. Technical Report, University of Toronto.

[77] Neal R. (2000), Markov Chain sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics,* Vol.9 No.2.

[78] Ng CM, Loyet KM, Lyer S, Fielder PJ, Deng R (2014), Modeling approach to investigate the effect of neonatal Fc receptor binding affinity and anti-therapeutic antibody on the pharmacokinetic of humanized monoclonal anti-tumor necrosis factor-a lg G antiboday in cynomolgus monkey. *Eur J Pharma Sci.* 51:51-8.

[79] Ng CM, Bai S, Takimoto CH, Tang MT, Tolcher AW (2010), Mechanism-based receptor-binding model to describe the pharmacokinetic and pharmacodynamic of an anti-51 integrin monoclonal antibody (volociximab) in cancer patients. *Cancer Chemother Pharmacol.* 65(2):207-17

[80] Roberts GO. and Rosenthal JS. (2007), Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.* 44(2):458-475

[81] Roberts GO, Rosenthal JS. and Schwartz PO.(1998) Convergence properties of perturbed Markov chains. *J. Appl. Probab.* 35(1):1-11

[82] Roberts GO, Papaspiliopoulos O. and Dellaportas P. (2004), Bayesian inference for non-Gaussian Orstein-Uhlenbeck stochastic volatility processes *J. Royal Statist. Soc. B* 66:369-933.

[83] Robert CP. and Casella G. (2004), Monte Carlo Statistical Methods. Springer, second edition.

[84] Rocchetti M, Del B.F., Germani M., Fiorentini F., Poggesi EP., Magni P.G. and Nicolao D. (2009), Testing additivity of anticancer agents in pre-clinical studies: a PK/PD modelling approach , *European Journal of Cancer*, 45, 3336-3346.

[85] Rosner G.L. and Muller P. (1994) Pharmacokinetic/Pharmacodynamic Analysis of Hematologic Profiles, *Journal of of Pharmacokinetics and Biopharmaceutics*, 22, 499524.

[86] Roosmarijn C., Chiara P., Elke K., Meindert D., Karel A., and Knibbe C. (2011), The role of population PKPD modelling in paediatric clinical research, *Eur J Clin Pharmacol*. 67(Suppl 1): 516.

[87] Olsen C.K., Brennum L.T. and Kreilgaard M. (2007). PK/PD modelling as a tool for prediction of therapeutic effective plasma levels of sertindole and other antipsychotics. *European Neuropsychopharmacology,* 17.

[88] Putter H., Keisterkamp S.H., Lange J.M. and Wolf F.d. (2002). A bayesian approach to parameter estimation in hiv dynamical models. *Statistics in Medicine*, 21, 2199-2214.

[89] Pasarica C. and Gelman A. (2010) Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, 20, 343-364

[90] Sudderth E.B. and Jordan M.I. (2009) Shared segmentation of natural scenes using dependent Pitman-Yor processes. *In Neural Information Processing Systems* 22.

[91] Saksman E. and Vihola M. (2008) On the Ergodicity of the Adaptive Metropolis Algorithms on Unbounded Domains. Preprint

[92] Sheiner and Beal (1980), Evaluation of methods for estimating population pharmacokinetic parameters. I Michaelis-Menten model: routine clinical pharmacokinetic data. *J pharmacokinet Biopham*, 8(6):553-71

[93] Sheiner and Beal (1981). Evaluation of methods for estimating population pharmacokinetic parameters. II Michaelis-Menten model: routine clinical pharmacokinetic data. *J pharmacokinet Biopham*, 9(5):635-51.

[94] Sheiner and Beal (1983), Evaluation of methods for estimating population pharmacokinetic parameters. III Michaelis-Menten model: routine clinical pharmacokinetic data. *J pharmacokinet Biopham*, 11(3):303-19.

[95] Kristoffer Sahlin (2011) Estiamating convergence of Markov chain Monte Carlo simulations. (Thesis manuscript)

[96] Smith B.J. (2000). Bayesian Output Analysis Program (BOA)Version 0.5.0 User Manual, Department of Biostatistcs, University of Iowa College of Public Health.

[97] Wang Y., Eskridge K. and Zhang S. (2008) Semi-parametric mixed-effects analysis of PK/PD models using differential equations *J Pharmacokinet Pharmacodyn*, 35:443-463

[98] Teorell T. (1937a) Kinetics of distribution of substances administered to the body I. The extravascular modes of administration. *Archs. int. Pharmacodyn. Ther*, 57:205-255.

[99] Teorell T. (1937b) Kinetics of distribution of substances administered to the body II. The intravascular modes of administration. *Archs. int. Pharmacodyn. Ther.*, 57: 226-240.

[100] Teh YW. (2006b) A Hierarchical Bayesian Language model based on Pitman-Yor processes. In ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 985-992, Morristown, NJ, USA. Association for Computational Linguistics.

[101] Teh YW. (2010) Dirichlet Process *Encyclopedia of Machine Learning*, Springer.

[102] Yamazaki S., Lam J., Zou H., Wang H., Smeal T. and Vicini P. (2014). Mechanistic understanding of translational pharmacokinetic-pharmacodynamic relationships in nonclinical tumor models: A case study of orally available novel inhibitors of anaplastic lymphoma Kinase. *Drug Metab. Disposo.*, 43(1):54-62.

[103] Yang C. (2008), Recurrent and Ergodic Properties of Adaptive MCMC, *Techinical Report in Department of Statistics at the University of Toronto* No.0801

[104] Wolfinger RD., (1993) Laplace's approximation for nonlinear mixed models, *Biometrika*, 80, 791-795.

[105] Wagner J.G. (1975), Do you need a pharmacokinetic model, and if so, which one. *J. Pharmacokinet. Biopharmaceut.*. 3457-478

[106] Wakefield, J.C. (1996), The bayesian analysis to population pharmacokinetic models. *Journal of the American Statistical Association*, 91, 62-75.

[107] Turchin VF. (1971), On the Computation of Multidimensional Integrals by the Monte Carlo Method. *Theory of Probability and its Applications*, 16(4) pp.720-724.

[108] Walker S. and Damien P. (1998), Sampling methods for Bayesian nonparametric inference involving stochastic processes, *Practical Nonparametric and Semiparametric Bayesian Statistics Lecture Notes in Statistics*, Volume 133, pp 243-254

[109] Ishwaran H. and James LF. (2002). Gibbs sampling methods for stick-breaking priors, *Journal of the Americn Statistical Association* 96 (453):161-173.

[110] Vonesh EF. and Carter RL. (1992), Mixed effects nonlinear regression for unbalanced repeated measures. *Biometrics* 48, 1-18.

[111] Zhang L, Beal SL, Sheiner LB. (2003) Simultaneous vs. sequential analysis for population PK/PD data I: best-case performance. *J. Pharmacokinet Pharmacodyn.* 30 (6):387-404.