

PENALIZED SPLINE MODELING OF THE *EX-VIVO* ASSAYS DOSE-  
RESPONSE CURVES AND THE HIV-INFECTED PATIENTS'  
BODYWEIGHT CHANGE

Samiha Sarwat

Submitted to the faculty of the University Graduate School  
in partial fulfillment of the requirements  
for the degree  
Doctor of Philosophy  
in the Department of Biostatistics  
Indiana University  
September 2015

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

---

Jaroslav Harezlak, PhD, Chair

---

Constantin T. Yiannoutsos, PhD

Doctoral Committee

---

Xiaochun Li, PhD

June 5, 2015

---

Kara K. Wools-Kaloustian, MD, MS

© 2015

Samiha Sarwat

## DEDICATION

To My Grandmother

## ACKNOWLEDGEMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Dr. Jaroslaw Harezlak, my advisor for his wonderful guidance as well as the enormous amount of hours that he spent on thinking through the projects and revising the writings. I am also very grateful to Dr. Constantin T. Yiannoutsos, my research committee advisor, for his willingness and precious time to guide and review my research work. Special thanks to Dr. Xiaochun Li and Dr. Kara K. Wools-Kaloustian, MD, for agreeing to serve on my committee and their careful and critical reading of this dissertation.

I would like to acknowledge and thank the department of the Biostatistics and the department of Mathematics for creating this PhD program and providing friendly academic environment. I also acknowledge the faculty, the staff and my fellow graduate student for their various supports during my graduate study.

Last but not least I would like to thank my family, my husband Nayan and my son Ian for their unconditional support, constant source of love, patience, concern, support and strength during all these years. My special thanks to my mother, Soheli and my cousin, Tushi for their continuous encouragement during all these years. I wish to express my heart-felt gratitude to my aunties, my cousins and friends, who have aided and encouraged me throughout this effort.

Samaha Sarwat

PENALIZED SPLINE MODELING OF THE EX-VIVO ASSAYS DOSE-  
RESPONSE CURVES AND THE HIV-INFECTED PATIENTS' BODYWEIGHT  
CHANGE

A semi-parametric approach incorporates parametric and nonparametric functions in the model and is very useful in situations when a fully parametric model is inadequate. The objective of this dissertation is to extend statistical methodology employing the semi-parametric modeling approach to analyze data in health science research areas. This dissertation has three parts. The first part discusses the modeling of the dose-response relationship with correlated data by introducing overall drug effects in addition to the deviation of each subject-specific curve from the population average. Here, a penalized spline regression method that allows modeling of the smooth dose-response relationship is applied to data in studies monitoring malaria drug resistance through the ex-vivo assays. The second part of the dissertation extends the SiZer map, which is an exploratory and a powerful visualization tool, to detect underlying significant features (increase, decrease, or no change) of the curve at various smoothing levels. Here, Penalized Spline Significant Zero Crossings of Derivatives (PS-SiZer), using a penalized spline regression, is introduced to investigate significant features in correlated data arising from longitudinal settings. The third part of the dissertation applies the proposed PS-SiZer methodology to analyze HIV data. The durability of significant weight change over a period is explored from the PS-SiZer visualization. PS-SiZer is a graphical tool for exploring structures in curves by mapping areas where rate of change is significantly increasing, decreasing, or does not change. PS-SiZer maps provide information about the significant rate of weight change that occurs in two ART regimens at various level of smoothing. A penalized spline

regression model at an optimum smoothing level is applied to obtain an estimated first-time point where weight no longer increases for different treatment regimens.

Jaroslav Harezlak, Ph.D., Chair

## TABLE OF CONTENTS

CHAPTER 1.	INTRODUCTION.....	1
	Overview of semi-parametric regression method: penalized spline regression.....	1
1.1	Penalized Spline Regression modeling to Analyze Dose-Response Functions and its Application to Monitoring Malaria Drug Resistance in Drug Assays .....	2
1.2	Penalized Spline Significant Zero Crossings of Derivatives (PS-SiZer): A Visual Tool to Investigate Significant Features in Longitudinal data .....	2
1.3	Application of PS-SiZer map to investigate significant features of the rate of change of body-weight profile for HIV infected patients in IeDEA study.....	3
CHAPTER 2.	AN OVERVIEW OF THE SEMI-PARAMTERIC REGRESSION METHOD: PENALIZED REGRESSION SPLINE .....	5
2.1	Semi-parametric regression analysis.....	5
2.2	Penalized spline regression in longitudinal studies .....	12
CHAPTER 3.	PENALIZED REGRESSION SPLINE MODELING OF DOSE-RESPONSE FUNCTIONS AND APPLICATION TO MONITORING MALARIA DRUG RESISTANCE IN DRUG ASSAYS .....	14
	Summary.....	14
3.1	Introduction.....	15
3.2	Motivating example – monitoring malaria drug resistance .....	21
3.3	Dose-response modeling and semi-parametric approach .....	22
3.3.1	PSDR-PS-POP model for overall drug and subject-specific effect.....	24
3.3.2	PSDR-PS- model for individual subject curves:.....	28
3.3.3	Algorithm to calculate relative $IC_{50}$ from the PS-POP and PS model.....	30



3.4	Simulation studies .....	32
3.5	PSDR model: application to malaria data .....	36
3.6	Conclusion and Discussion .....	39
	REFERENCES .....	41
CHAPTER 4. PS-SIZER: A VISUAL TOOL TO INVESTIGATE		
	SIGNIFICANT FEATURES IN LONGITUDINAL DATA.....	44
4.1	Summary .....	44
4.2	Introduction.....	44
4.3	SiZer method.....	51
4.3.1	LL-SiZer .....	51
4.3.2	SS-SiZer.....	53
4.4	PS-SiZer method.....	54
4.4.1	Inference .....	58
4.4.2	Estimate and variability bands of the derivatives .....	59
4.4.3	Confidence band .....	59
4.4.4	Construction of color coded PS-SiZer map .....	61
4.5	Simulation study .....	61
4.5.1	Simulation study-one .....	63
4.5.2	Simulation study-two .....	67
4.6	Application of PS-SiZer map.....	72
4.6.1	Results from PS-SiZer .....	72

4.7	Discussion and conclusion.....	74
	REFERENCES .....	76
CHAPTER 5. AN APPLICATION OF PS-SIZER MAP TO INVESTIGATE		
SIGNIFICANT FEATURES OF THE RATE OF CHANGE OF BODY-WEIGHT		
PROFILE FOR HIV INFECTED PATIENTS IN IEDEA STUDY.....		
		79
	Abstract .....	79
5.1	Introduction.....	80
5.2	Methods.....	84
5.2.1	Population .....	84
5.2.2	Statistical analysis .....	85
5.2.3	PS-SiZer Maps .....	85
5.2.4	Algorithm to detect first time point (week) at which weight gain stops	
	increasing at an optimum smoothing level .....	87
5.3	Results.....	89
5.3.1	Baseline characteristics.....	89
5.3.2	Exploratory data analysis.....	92
5.3.3	PS-SiZer maps and durability of weight gain at optimum smoothing.....	93
5.4	Discussion and Conclusions .....	98
5.5	Limitations .....	99
5.6	Strengths .....	100
	REFERENCES .....	106
CURRICULUM VITAE		

## CHAPTER 1. INTRODUCTION

The objective of this dissertation is to extend statistical methodology employing the semi-parametric modeling approach to analyze longitudinal data in health science research areas. Various parametric and non-parametric models and statistical tools has been developed to analyze longitudinal data. The parametric models assume a predefined parametric relationship between the response variables and its covariates which may lead to modeling biases when such relationship is not known. On the other hand non-parametric models are too flexible to make concise conclusion compared to parametric models. Semi-parametric models are good compromises with good features from both parametric and non-parametric model. Semi-parametric models are useful when the functional form of the parametric model is unknown or in a situation when a fully non-parametric model may not perform well.

Penalized spline regression models are popular semi-parametric statistical tools for curve fitting, because of their flexibility and computational efficiency, but not widely used by researchers in disciplines outside of statistical sciences. The goal of my dissertation papers is to extend statistical models for correlated observations arising from longitudinal settings by utilizing penalized spline regression approach, specifically, in the field of dose-response analysis and in the epidemiologic studies of Human Immunodeficiency Virus (HIV).

### Overview of semi-parametric regression method: penalized spline regression

In this chapter, a brief overview of semi-parametric regression is presented in a concise and modular fashion.

## 1.1 Penalized Spline Regression modeling to Analyze Dose-Response Functions and its Application to Monitoring Malaria Drug Resistance in Drug Assays

Dose-response assays describe the effect of changes in an organism growth caused by exposure to increasing drug concentration. The analysis of such experiments frequently relies on parametric sigmoidal (logistic) models. However, dose-response data often do not follow the pre-specified shape. Therefore, we need flexible modeling approaches. We propose an application of a penalized spline regression method that allows modeling of the smooth dose response relationship with correlated data. We call our model, Penalized Spline Dose-Response (PSDR) method. We use the PSDR method to analyze data arising in the studies monitoring malaria drug resistance through *ex vivo* assays. Our objectives of this research are: (i) to model dose-response relationship with correlated data by introducing overall drug effects in addition to the analysis based on each biological replicate. (ii) to estimate the quantities of interest, e.g. half-maximal inhibitory concentration ( $IC_{50}$ ) and obtain their properties (standard errors – SE and confidence intervals - CI). (iii) to develop a user friendly R-function for the analysis of the PSDR models.

## 1.2 Penalized Spline Significant Zero Crossings of Derivatives (PS-SiZer): A Visual Tool to Investigate Significant Features in Longitudinal data

We propose an extension of the Significant Zero Crossings of Derivatives (SiZer) as an exploratory graphical tool to analyze longitudinal data. The standard implementation of SiZer is based on the local linear smoother with kernel-type smoothing method for curve estimation problems. In longitudinal studies, data are often correlated and it is necessary to account for the within-subject correlation. In this work, we propose an extension of the

SiZer methodology for correlated observations arising from longitudinal settings by utilizing penalized spline regression model. The proposed approach is an extension of the SiZer map for correlated observations arising from longitudinal settings and enhancement of the SiZer analysis using computationally efficient smoothing method, penalized spline regression model. We apply our PS-SiZer methodology to analyze the differential pattern of weight change over time among the HIV patients, data from the International Epidemiologic Database to Evaluate AIDS (IeDEA) collaboration.

### 1.3 Application of PS-SiZer map to investigate significant features of the rate of change of body-weight profile for HIV infected patients in IeDEA study

This work involves standardized data collected on HIV-positive patients initiating antiretroviral therapy (ART) in five regions of the International Epidemiologic Databases to Evaluate AIDS (IeDEA) collaboration. The key objective is to understand the pattern of body-weight change in HIV patients initiating stavudine (d4T) containing first-line regimens versus non-d4T-containing regimens. PS-SiZer is a unique visualization tool to investigate significant features which can handle longitudinally collected data, such as body-weight change in HIV patients. PS-SiZer map used to explore the structure in curves by mapping areas where weight change is significantly increasing, decreasing or does not change. The PS-SiZer map together with a fitted smoothed curve at an optimum level of smoothing provided valuable insight and used for statistical comparison of the durability of weight gain in patients received ART regimens containing and not containing d4T.

This dissertation is organized as follows. In Chapter 2, we present an overview of semi-parametric regression models. In Chapter 3, we present penalized spline regression modeling to analyze dose-response functions and its application to monitoring malaria drug

resistance in *ex-vivo* drug assays. We present, PS-SiZer: a visual tool to investigate significant features in longitudinal data in Chapter 4. In Chapter 5, we present application of PS-SiZer to investigate significant features of the body-weight profile for HIV infected patients in IeDEA study.

## CHAPTER 2. AN OVERVIEW OF THE SEMI-PARAMTERIC REGRESSION

### METHOD: PENALIZED REGRESSION SPLINE

#### 2.1 Semi-parametric regression analysis

Semi-parametric regression for mean modeling for independent data have been well developed over more than two decades (Green & Silverman, 1994). Such regression models combine parametric functions of a subset of the covariates and non-parametric functions of other covariates to model the mean of an outcome variable. Semi-parametric models are useful when the functional form of the parametric model is unknown or in a situation when a fully non-parametric model may not perform well. Non-parametric or semi-parametric regression methods can be broadly classified into kernel methods and splines. A book authored by Wand and Jones (1995) is an excellent source of kernel methodology which was expanded upon by Fan and Gijbel (1996) based on the local likelihood approach. Spline or smoothing spline is another attractive semi-parametric technique that has gained popularity in the last 20 years. Spline techniques include smoothing spline [Green & Silverman (1994); Wahba (1990)], regression spline (Stone, Hansen, Kooperberg, & Truong, 1997) and penalized spline [Eilers & Marx (1996); Ruppert, Wand, & Carroll (2003)]. Spline techniques offer more flexibility than traditional parametric polynomial regression for fitting non-linear and non-polynomial relationship. Some definitions and related literature are provided in the following subsections.

#### Smoothing splines

The term 'spline' describes the process of fitting a piecewise polynomial function to data points. A smoothing spline estimates the regression function with all the observed covariate values used as knots using a piecewise polynomial function. A knot is defined as

the point at which piecewise polynomials are joined together. For example, the most commonly used smoothing spline is the natural cubic smoothing spline which assumes the piecewise cubic function as the smoothing function and is continuous and twice differentiable at the knots.

### Regression spline.

A regression spline is a spline that considers a small number of knots and proceeds with a parametric regression using bases. For example,  $\beta_0 + \beta_1 x$  is a linear combination of the basis function 1 and  $x$ . Thus,  $\{1, x\}$  is a basis for the vector space of all linear polynomials in  $x$ . Basis functions, such as B-spline and truncated polynomial basis, and radial basis are some examples of basis functions used in practice.

In a regression spline, one needs to select the number and location of the knots as well as a set of basis function. Fitting of such a model tends to depend quite strongly on the number and locations chosen for the knots. Smooth curves can be parametrically modeled using a regression spline basis. However, sometimes a low-dimensional basis is difficult to select. An alternative to controlling smoothness is to select a high-dimensional basis but, then, penalizing the estimated coefficients by adding a ‘wiggleness’ penalty to the least squares fitting objective. This approach leads to a simple and flexible spline based regression model known as a penalized spline regression.

### Penalized spline regression

Eilers and Marx (1996) proposed the technique of penalized splines, a method of fitting a smoothing spline using penalties to constrain the roughness of the fit. Penalized spline regression is a combination of a regression spline and smoothing spline (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008). Moreover, penalized spline has



close ties with ridge regression and mixed models, ties that were discovered by researchers working on smoothing splines. These ties allow techniques from mixed models, for example, Restricted Maximum Likelihood estimation (REML), likelihood ratio tests, to be added to penalized spline methodology.

The general definition of penalized spline regression as described in Ruppert et al. (2003) required two basic choices:

- (1) The spline model – that is, the degree and knot locations and whether to impose constraints such as a boundary constraints and
- (2) The penalty – or, more explicitly, the form of the penalty up to a nonnegative smoothing parameter.

Once these two choices have been made, two secondary choices follow:

- (3) The basis function – for example, truncated power function or  $B$ -spline to represent the model matrix and
- (4) The basis function used in the computation.

The later choices do not affect the fitted curve with exception of the effects of numerical error. Once the penalty and the basis function have been determined, then the penalty matrix is automatically determined.

The various types of penalized splines can be tied together with a broader concept. Given a scatter plot data  $(x_i, y_i), i = 1, \dots, n$ , let  $B(x) = [B_1(x), \dots, B_N(x)]^T$  be the vector of spline basis functions, so that the  $i^{th}$  row of  $X$  is  $B_i(x)^T$ . The general definition of a penalized spline is  $\hat{\beta}^T B(x)$ , where  $\hat{\beta}$  is the minimizer of

$$\sum_{i=1}^n \{y_i - \beta^T B(x)\}^2 + \lambda \beta^T D \beta$$

for some symmetric positive semidefinite matrix  $D$  and the scalar  $\lambda > 0$ .

Two basis functions represented by truncated power functions or  $B$ -splines to characterize the smoothing model are briefly discussed in the following sub-sections.

### Penalized spline model using truncated basis function

Ruppert, Wand & Carroll (2003) simplified the spline mathematics by using a truncated line basis function. The truncated line basis function is represented as,

$$(x - k_m)_+ = \begin{cases} (x - k_m), & x > k_m \\ 0, & x \leq k_m \end{cases}$$

$k_m$  is the  $m^{\text{th}}$  knot. The spline model can be written as,

$$y_i = \beta_0 + \beta_1(x_i) + \sum_{k=1}^K \beta_{1k} (x_i - k_k)_+ + \varepsilon_i,$$

where  $\beta = [\beta_0 \ \beta_1 \ \beta_{11} \ \dots \ \beta_{1K}]^T$  are the coefficients of the polynomial functions and truncated line functions. The design matrix is represented as,

$$X = [1 \ x_i \ (x_i - k_k)_+]_{1 \leq i \leq n}$$

The Penalized spline fitting criteria is

$$\text{minimize } \|y - X\beta\|^2 + \lambda \beta^T D \beta$$

where  $D$  is a symmetric positive semi-definite penalty matrix such that  $D = \text{diag}(0,0,1, \dots, 1)$ , and  $\lambda$  is a smoothing parameter which controls the amount of smoothing. Then the solution for the regression coefficients is  $\hat{\beta} = (X^T X + \lambda D)^{-1}$  and the fitted value for a penalized regression spline are then given by,

$$\hat{y} = X(X^T X + \lambda D)^{-1} X^T y$$

### Mixed model representation of penalized spline using truncated basis function

The book “Semiparametric Regression” authored by Ruppert, Wand, & Carroll (2003) provided details connection of the solution to the penalized spline criterion as a

BLUP in a mixed model framework. This is very useful because it allows smoothing to be done using mixed model methodology and existing statistical software. The connection between spline and mixed model arises from the similarity of the penalized spline fitting criterion to the minimization problem that yields the mixed model equations and solutions. Ruppert, Wand, & Carroll's (2003) approach makes splines much more accessible since it relies on the well understood mixed model theory.

This connection is made explicit in the Ngo & Wand (2004) paper. The authors provided S-PLUS and SAS code that illustrates the use of mixed model software to do smoothing for several penalized spline models.

It is useful to rewrite the function  $f(\cdot)$  as a mixed effect model. Let the linear spline model for  $f$  be

$$f(x_i) = \beta_0 + \beta_1(x_i) + \sum_{k=1}^K u_k (x_i - k_k)_+$$

Let  $\beta = [\beta_0 \ \beta_1]^T$  and  $u = [u_1 \ u_2 \ \dots \ u_K]^T$ . The design matrices are defined as:

$X = [1 \ x_i]_{1 \leq i \leq n}$  and  $Z = [(x_i - k_k)_+]_{1 \leq i \leq n, 1 \leq k \leq K}$ . We can rewrite the objective function dividing by  $\sigma_\varepsilon^2$  as:

$$\frac{1}{\sigma_\varepsilon^2} \|y - X\beta - Zu\|^2 + \frac{\lambda}{\sigma_\varepsilon^2} \|u\|^2$$

This objective function is equivalent to the BLUP criteria by treating  $u$  as a set of random coefficients with  $Cov(u) = \sigma_u^2 I$  and  $\sigma_u^2 = \frac{\sigma_\varepsilon^2}{\lambda}$ . Putting all this together, the mixed model representation of the penalized spline regression is obtained as below:

$$y = X\beta + Zu + \varepsilon, \quad Cov \begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sigma_u^2 I & 0 \\ 0 & \sigma_\varepsilon^2 I \end{bmatrix}$$

The estimation of the fitted smoothing model is obtained using existing mixed model software. The vector of parameters  $\beta$  and the random coefficient vector  $u$  can be determined using best prediction. Conditional on the REML estimates of the smoothing parameters and other variance components, the estimates of the splines are simply empirical best linear unbiased predictions (EBLUPs).

### Penalized spline regression using P-spline bases

P-spline (Eilers & Marx, 1996) is defined as a combination of B-spline bases and difference penalties. P-splines are smoothing splines based on B-spline basis function. P-splines expand the method of using a penalty to control the smoothing, where the penalty is based on the difference of the coefficients of adjacent B-splines. Here, B-splines are constructed from polynomial pieces and joined at equally spaced knots. Once the knots are defined, B-splines are computed recursively for any defined degree of the polynomial (De Boor, 1978). Let  $B_m(x_{ij}; p)$  denote B spline basis of degree  $p$  with  $m'$  equal intervals of  $m' + 1$  knots. Hence the number of B-spline in the regression is  $M = m' + p$ . The algorithm to compute B-spline basis of any degree was detailed in the book by De Boor (1978).

The recursive formula adopted by Eilers and Marx (1996) is shown as follow:

$$B_j(x; p) = \frac{x-x_j}{x_{j+p+1}-x_j} B_j(x; p-1) + \frac{x_{j+p+2}-x}{x_{j+p+2}-x_{j+1}} B_{j+1}(x; p-1)$$

$$B_j(x; -1) = \begin{cases} 1, & x_j \leq x < x_{j+1} \\ 0, & \text{otherwise} \end{cases}$$

where  $B_j(x; p)$  is the basis function evaluated at  $x_j$ , and  $p$  is the order of the basis function being calculated. Consider the regression of  $n$  data points  $(x_i, y_i)$  on a set of  $M$  B-Splines  $B_j(\cdot)$ . The least square objective function to minimize is then taking a form as below:

$$\sum_{i=1}^n \left\{ y_i - \sum_{j=1}^M a_j B_j(x_i) \right\}^2$$

where  $a_j$  is the vector of coefficients. Eilers and Marx (1996) proposed a finite difference penalty to construct the P-Spline model for example, difference penalty with order  $d$  can be written as:  $a^T D_d^T D_d a$ , where  $D_d$  is the  $d^{\text{th}}$  order finite difference penalty matrix, and  $D_d a$  is the vector of  $d^{\text{th}}$  difference of  $a$ .

Hence, the penalized objective functions take the form below:

$$\|y - \mathbf{B}a\|^2 + \lambda \|D_d a\|^2$$

The parameter  $\lambda$  is the smoothing parameter to control the wiggleness of the fit. This approach reduces the dimensionality of the problem to  $M$ , the number of B-splines, instead of  $n$ , the total number of observations.

#### Mixed model representation of P-spline

Similar to the penalized spline with truncated polynomial basis, P-spline can be represented in a mixed model framework. The minimization problem is handled using the mixed model framework by treating the smoothing component as a random component of the mixed model (Currie & Durban, 2001). Hence, the mixed model representation is:

$Y = X\beta + Za + \varepsilon$ , where,  $X$  is the fixed effect part of the model with  $\beta$  to be estimated, and  $Z$  is the model matrix from smoothing component with  $a \sim N(0, \sigma_\varepsilon^2 (\lambda D_d^T D_d)^{-})$  and  $\varepsilon$  the vector of error variance with  $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ . Here  $a$  is treated as a random vector but  $a$  has an improper distribution. The improper distribution for  $a$  does not fit easily into standard linear mixed modeling approaches (Pinheiro & Bates, 2000). Some re-parameterization is needed. So that the new parameters are divided into a set with a proper

distribution, to be treated as random effects, and a set with improper uniform distribution, to be treated as fixed effects (Wood S. N., 2006b).

## 2.2 Penalized spline regression in longitudinal studies

In the last 15 years, significant developments have taken place in semi-parametric regression methods for longitudinal data. In longitudinal studies, measurements are frequently collected for several subjects, and data are subject to within-subject variability. Extension of kernel and spline smoothing methods for longitudinal data is challenging due to the presence of within-subject correlation among repeated measurements over time.

There have been several substantial research studies done in developing non-parametric estimation procedures under the setting of clustered or longitudinal data. Lin and Carroll (2001) proposed kernel Generalized Estimation Equations (GEE) and showed that the kernel GEE works the best without incorporating within-subject correlation. Classical local-likelihood based kernel methods fail to effectively account for the within-subject correlation. Wang (2003) proposed Seemingly Unrelated (SUR) kernel estimator using an iterative algorithm. Linton et al. (2004) proposed an estimator using different pseudo-observations that is more efficient than kernel GEE but less efficient than the SUR kernel in presence of within-subject correlation.

A smoothing spline estimates the non-parametric regression function using piecewise polynomial function with all the observed covariate values used as knots, where smoothness constraints are assumed at the knots. The presence of the within-subject correlation among repeated measurements over time presents a major challenge in spline smoothing for longitudinal data. Extension of spline smoothing to longitudinal data requires explicitly accounting for the within-subject correlation in the likelihood function.

Lin et al. (2004) showed smoothing spline estimator is asymptotically equivalent to the SUR kernel estimator. Smoothing spline method provides several attractive features compared to kernel smoothing. A smoothing spline estimator has a close connection with the linear mixed model and can be obtained by fitting the linear mixed model. One can treat a smoothing parameter as an extra variance component in addition to variance components in the model and can simultaneously estimate by using restricted maximum likelihood (RELM) under the linear mixed model.

Since a smoothing spline uses all data points as knots, large data computation can become cumbersome. A penalized spline regression detailed by Eilers & Marx (1996) and Ruppert, Wand, & Carroll (2003), among others is an attractive alternative. Penalized spline regression uses a moderate number of knots, and the penalty approach controls the wiggleness of the smoothing function. The estimates of the regression coefficients are obtained via the penalized likelihood approach.

The mixed model representation of the penalized spline regression model allows for mixed model estimation techniques, and computation can be done by using existing statistical software. The mixed model representation of penalized splines allows for a seamless fusion between parametric mixed models and smoothing. For large data, computation is less expensive than smoothing spline and kernel based methods.

CHAPTER 3. PENALIZED REGRESSION SPLINE MODELING OF DOSE-  
RESPONSE FUNCTIONS AND APPLICATION TO MONITORING MALARIA  
DRUG RESISTANCE IN DRUG ASSAYS

Summary

Dose-response assays describe the effect of changes in the growth of an organism caused by exposure to increasing drug concentration. The analysis of such experiments frequently relies on parametric sigmoidal (logistic) models. However, dose-response data often do not follow a pre-specified shape, and more flexible modeling approaches are necessary. We propose a penalized spline dose-response (PSDR) method, a particular semi-parametric case of penalized regression splines that allow modeling of the smooth dose-response relationship with correlated data via linear mixed model representation. The PSDR method preserves the hierarchy of the technical and biological replicates while letting the data guide the model estimates. We used the PSDR method to analyze data arising from a study that monitored malaria drug resistance through *ex vivo* assays and obtained the quantities of interest (e.g. half-maximal inhibitory concentration ( $IC_{50}$ ) and their associated properties).

KEY WORDS: Dose response; Penalized spline regression; Semi-parametric.



### 3.1 Introduction

Dose-response curves are used in several stages of drug development. First, they may be used to identify drug targets in drug screening assays. Second, in Phase II studies, they are used to support decisions about effective and safe doses of a drug. Finally, they may be used to monitor development of drug resistance through drug sensitivity assays. Various study designs and statistical analysis methods have been developed to analyze dose-response data, such as model-based approaches that assume a functional relationship between the dose and response. Model-based approaches can be parametric regression with several functional forms, including the popular non-linear sigmoidal curves. One special case is the log-logistic curve; these curves with four or five parameters have direct biological interpretation which has made them increasingly popular.

However, it may be difficult to find a parametric model that fits data from the majority of dose-response experiments which leads to an inadequate fit. While analyzing *ex vivo* drug sensitivity assays used to monitor decreases in drug sensitivity by the malaria parasite, we found several cases in which the observed relationships between drug and response did not fit a single parametric model. Motivated by the necessity to obtain adequate estimates of drug sensitivity in these assays, we propose a semi-parametric approach to estimate dose-response curves and summary measures of drug sensitivity that could be used in any dose-response analysis, particularly in drug sensitivity assays. The proposed approach can also be extended to estimate dose-response curves when searching for the drug targets or effective and safe doses of a drug.

Traditionally, statistical analysis of dose-finding studies were developed using the multiple comparison procedure (MCP) (Bretz, Pinheiro, & Branson, 2005). The primary

goal of MCP is often to identify the minimum effective dose (MED) which considers dose as a qualitative factor and makes very few assumptions about the underlying dose- response model. In addition, the parametric model-based approach gained popularity in recent dose-response literature. The model-based approach assumes a functional relationship between the response and dose, according to a pre-specified parametric model. The fitted model is then used to represent the dose–response relationship and, subsequently, estimate an adequate dose to achieve a desired response.

Steenland and Deddens (2004) discussed model-based methods using a categorical approach, regression-splines, and simple parametric models to evaluate dose-response relationships with real data examples and mentioned their corresponding estimation and prediction issues. In their paper, Steenland and Deddens (2004) summarized their findings with examples that the best fitting model such as a pre-specified parametric model might not necessarily be the best model for risk assessment in dose-response studies. Further, Bretz et al. (2005) discussed a strategy to combine two major dose-response analysis approaches: MCP and model-based analysis. Their proposed approach assumed that there are several parametric models available and that the multiple comparison technique would be used to choose the best model for the underlying dose-response data. Several problems with this method can be noted. For example, a set of suitable parametric models needs to exist. There is no consideration for the correlation, repeated measurements, or the sensitivity of the choice of initial values for the standardized models. Another possible solution to fit dose-response data with a non-parametric method such as kernel regression or local linear regression which was first introduced by Cleveland (1979) followed by Fan and Gijbels (1992).

To quantify non-linear exposure-response associations, the most common models used in practice are Emax, log-logistic, exponential decay and quadratic (Bretz, Pinheiro, & Branson, 2005). The Emax model represents the percentage of the maximum change from the basal effect (dose=0) associated with dose  $d$ . The exponential decay with three parameters (EXP3) model intends to capture a sub-linear or a convex dose-response relationship. The log-logistic with four parameters (LL4) corresponds to four components: a basal effect, maximum effect from basal, the dose that gives half of the maximum change in effect and finally a parameter controlling the rate of change with dose in the effect. A quadratic (QUAD) model intends to capture a possible non-monotonic dose-response curve in either a concave or a convex shape. The functional form and the pre-specified shape of the 3-candidate models (LL4, EXP3, and QUAD) in dose-response analysis are presented in Table 3-1. On the other hand, to avoid parametric constraints on the shape of the exposure-response curve, a variety of smoothing techniques have been applied by epidemiologists (Greenland, 1995), (Govindarajulu, Malloy, Ganguli, Spiegelman, & Eisen, 2009). Some recent dose-response analysis is presented by fitting quadratic and cubic functions such as the restricted cubic splines (Desquilbet & Mariotti, 2010) or applying fractional polynomials (Govindarajulu, Malloy, Ganguli, Spiegelman, & Eisen, 2009). Desquilbet and Mariotti (2010) discussed restricted cubic spline (RCS) functions as a useful tool to analyze dose-response relationships where continuous exposure and the response have a non-linear association. But, the basic disadvantage of RCS functions is that the shape of the dose-response association generally depends on the location and the number of knots.

As discussed in the literatures above, no single parametric model may be appropriate for all subjects, and the widely used sigmoidal models such as LL4 or EXD3 do not provide an adequate fit for the data. To better understand this situation, consider the data from the malaria assay (described in Figure 3-1 and Section 3.2) of two subjects. For the first subject (Figure 3-1 (a)), comparison of observed and fitted curves suggest that none of the parametric models (LL4, EXD3, and QUAD) fit the observed data well. The poor fit of the parametric models suggest that a semi-parametric procedure could be more appropriate. In the case of the second subject (although a fitted LL4 curve seemed to overlap the observed data (Figure 3-1 (b)), the estimated  $IC_{50}$  was incorrectly estimated and is unrealistic with a negative estimated dose (-0.45; SE = 3.98). Even with the fitted EXD3 model, the estimated  $IC_{50}$  was high (1.78; SE = 0.39). Achieving a good model-fit but implausible target summary measurements, such as risk assessment in dose-response analysis, has been previously noted by Steenland & Deddens (2004), who showed various examples of data well fitted with parametric models (log-log, log-linear) yielding biologically implausible (rapid increase) results. In summary, we observed that estimates based on the standard sigmoidal model were sensitive to the specific choice of the parametric non-linear model. The estimated dose-response curves may be incompetent and fail to fit the observed data, and even if a non-linear parametric model fits the data well, the model may fail to estimate the  $IC_{50}$  correctly.

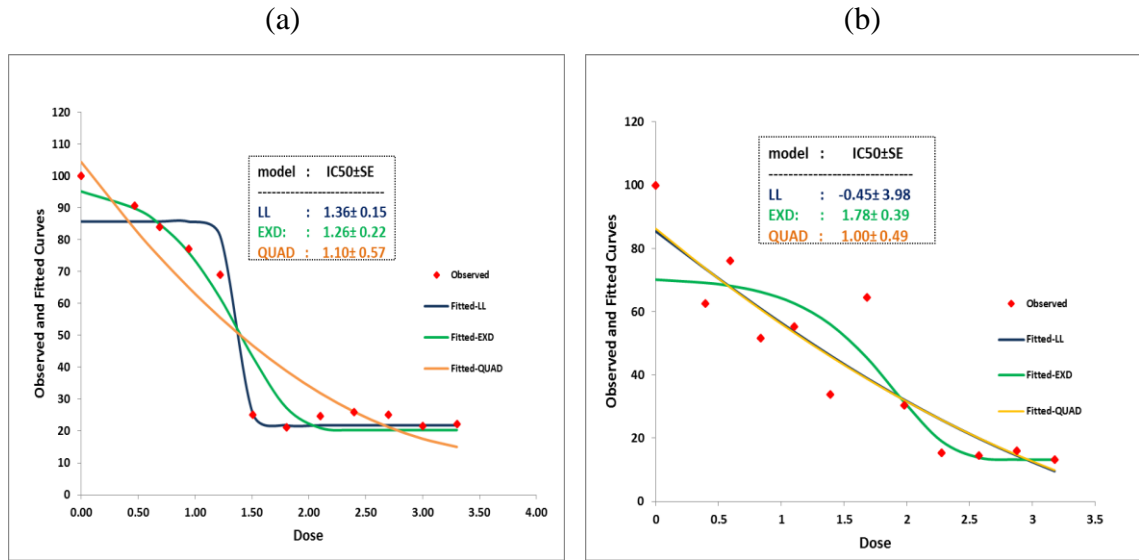


Figure 3-1 Fitted curves by LL4, EXD3 and QUAD parametric models and corresponding estimated  $IC_{50}$  with SD using malaria data. (a) Poor fit of the parametric models (b) parametric models fitted the data satisfactorily but the estimate of  $IC_{50}$  is unrealistic with large SD.  $IC_{50}$  values are in nM.

Although sigmoidal dose-response curves are biologically plausible to represent data from drug assays, we often encountered problems when analyzing several assays. For example, observed curves of more than 10% subjects clearly and visually did not fit the sigmoidal shape. Several mechanisms may contribute to obtaining curves with diverse shapes. For instance, subject parasites may present widely unanticipated variable sensitivity to the drug, and the doses used in the assay may cover only parts of the curve presented by the different subject parasite. Therefore, although the underlying true curve of dose-response is sigmoidal, only a fraction of this curve may be observed for some subjects. *Ex vivo* drug sensitivity assays are not only commonly used in malaria but also in other infectious diseases and cancer.

This paper focuses on the use of the penalized spline regression in the health sciences in the dose-response analysis. In our research, we applied a semi-parametric approach to remedy an inadequate parametric model by incorporating parametric and

nonparametric functions in the model. The book “Semiparametric Regression” authored by Ruppert, Wand, & Carroll (2003) is an excellent reference for nonparametric and semi-parametric regression models based upon penalized-splines. The book illustrates methodical approach of a mixed model representation with truncated polynomial bases, as almost immediate from the form of the basis and the penalty function. Eilers and Marx (1996) proposed a technique of penalized splines, a method of fitting a smoothing spline using knots and simple penalties. Ngo and Wand (2004) demonstrated a method using existing mixed model software by using the shrinkage property of mixed models instead of an externally defined penalty. Durb`an et.al (2005) proposed a penalized spline regression approach to model the deviation of each subject curve from the population average.

Our proposed semi-parametric model is called penalized spline dose-response (PSDR), an application of Durb`an et.al (2005) in the setting of dose-response analysis. This method preserves the hierarchy of the technical and biological replicates while letting the potentially correlated data guide the mean model estimates.

Our objectives for this undertaking were as follows (1) to allow the modeling of the dose-response relationship with correlated data by introducing overall drug effects in addition to the analysis based on each biological replicates (2) to estimate the quantities of interest (e.g. half maximal inhibitory concentration ( $IC_{50}$ ) and obtain their properties (standard deviation (SD), confidence limits (CL)), and (3) to develop a user-friendly R-function for the analysis of the dose-response relationship using the PSDR models.

This paper is organized as follows. In Section 3.2, we present data on monitoring drug sensitivity in malaria. Penalized regression splines method and its usefulness in dose-

response area are presented in section 3.3. In Section 3.4, simulation studies are presented to compare standard dose-response models with PSDR models. We apply our method to analyze a study from malaria data for screening compounds through an *ex vivo* experiment in Section 3.5, and concluding remarks are found in Section 3.6.

### 3.2 Motivating example – monitoring malaria drug resistance

Rising concerns with resistance to the whole malaria drug arsenal has led to widespread implementation of *ex vivo* assays to monitor sensitivity to several drugs in malaria point of point-of-care units in endemic regions of Africa, South America, and Asia. Upon seeking care to treat malaria where drug sensitivity is monitored, patients have venous blood drawn and subsequently receive treatment. *Ex vivo* assays are conducted with fresh blood taken directly from the patient and added to assay plates (generally 96 wells in 8 x 12 matrices) in which several sets of wells contain decreased concentrations of alternative drugs. The patient blood is left to interact with the drug in the plate for a pre-determined duration, and then parasite density is indirectly recorded using some type of marker, such as quantification of an antibody against the parasite or parasite DNA. A dose-response curve is fit for each patient and summarized through  $IC_{50}$  that is defined as the drug concentration causing fifty-percent inhibition of the desired activity.

In our research, we analyzed data from an *ex-vivo* 4',6-diamidino-2-phenylindole (DAPI) malaria drug assay including three drugs, Amodiaquine (AMQ), Chloroquine (CQ), and Mefloquine (MEF) and 12 doses, carried out in 106 subjects in a Senegalese clinic in 2008 (Ndiaye, et al., 2010). The arrangement of the collected data is presented in Table 3-1 for the drug AMQ as an illustration. The assay outcome was based on parasite load quantified through the fluorescence emitted by the DAPI dye that adhered to parasite

genetic material. The response of interest was the percentage of fluorescent decline (surrogate of percent decline in parasite density).

Table 3-1 An ex-vivo assays of Malaria drug resistance

Dr ug	Biological Replicates-Subjects	12 Doses (log 10)	Technical Replicates
A	106	0.00,0.47,0.69,0.95,1.22,1.51,1.80,2.10,2.40,2.70,3.00,3.30	2

### 3.3 Dose-response modeling and semi-parametric approach

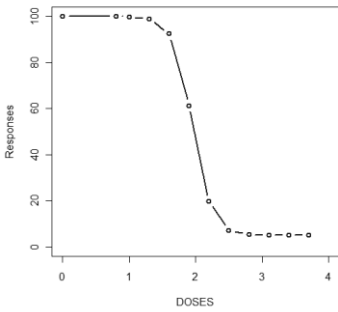
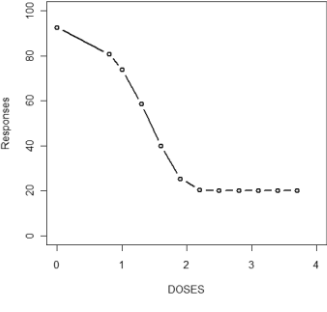
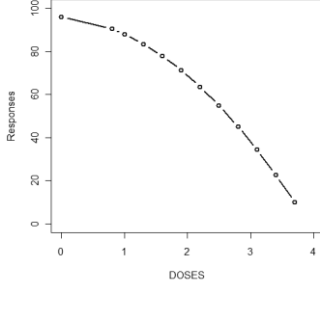
In practice, the functional form of the dose-response curve represented parametrically, such that the response of efficacy or safety variable is denoted by  $Y$ , which is observed for a given set of doses,  $d$ . The general framework is then defined as below:

$$Y_{ij} = \mu d_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \quad (3.1)$$

Where  $Y_{ij}$  = percent of parasitemia response of  $i^{\text{th}}$  subject at  $j^{\text{th}}$  dose,  $j = 1, \dots, n_i$  and  $i = 1, \dots, n$ . The mean response at dose  $d_i$  can be represented as  $\mu d_i = f(\cdot)$  for a dose-response model and may be a linear or nonlinear function of the model parameters.



Table 3-2 Selected candidate models used in the analysis of dose-response relationships

<b>Models*</b>		
<b>Functional representation</b>		
<b>Graphical representation</b>		
<b>Log- Logistic with 4- parameters (LL4).</b>	<b>Exponential decay with 3- parameters (EXD3).</b>	<b>Quadratic (QUAD).</b>
$f_{LL4}(x)**$ = $c$ + $\frac{(u-l)}{1 + \exp[b(\log(x) - \log(e))]}$	$f_{EXD3}(x)$ = $l + (u-l)(\exp(-x/e))$	$f_{Quad}(x)$ = $\beta_0 + \beta_1 x + \beta_2 x^2$
		
<p>Note: *Model fitted for individual subjects. ** l = lower limit, u = upper limit, b = slope, e=dose giving % of response, x=dose-level.</p>		

We propose to use penalized spline regression model to estimate dose-response curves and curve features and name it penalized spline dose response (PSDR) model.

The major motivation of this paper was to describe and account for the overall population or drug effect in the analysis of a dose-response relationship. We wanted to account for the overall drug effect as well as the deviation of each subject specific effect from the overall drug effect. Additionally, we wanted to consider the subject specific correlation in the dose-response model. In the following subsections we briefly present two PSDR models PS-POP and PS. The PS-POP model was developed to incorporate the population curve as

well as the deviation of each subject curve from the population average in the dose-response analysis and described in details in Section 3.3.1. PS-model was developed for individual subject-specific curve to follow the standard practice in dose-response analysis and described in Section 3.3.2.

### 3.3.1 PSDR-PS-POP model for overall drug and subject-specific effect

The PS-POP model is the application version of the semi-parametric modeling approach for dose-response settings originally proposed in Durban et.al (2005) for fitting subject-specific curves for longitudinal data.

Consider the dose-response standard model (equation 3.1), the penalized splines regression model version of the dose-response curve is formulated as

$$Y_{ij} = f_{drug}(x_{ij}) + g_i(x_{ij}) + \varepsilon_{ij}; \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \quad (3.2)$$

where  $Y_{ij}$  is the percent of parasitemia responses of  $i^{\text{th}}$  subject at dose  $x_{ij}$ , for  $i=1, \dots, n$  and  $j=1, \dots, n_i$ ,  $f_{drug}(x_{ij})$  is the population or overall drug effect, and  $g_i(x_{ij})$  is the deviation or departure of the  $i^{\text{th}}$  subject from the overall drug effect. So, there are two parts: the overall drug effect  $f_{drug}(x_{ij})$  and the subject-specific effect  $g_i(x_{ij})$ .

A smooth function is assigned to measure the overall drug effect,  $f_{drug}(x_{ij})$ . Similarly, a random smooth function is assigned to measure the subject specific effects  $g_i(x_{ij})$ . There are various approaches for modeling  $f_{drug}(x_{ij})$  and  $g_i(x_{ij})$  with associated penalties. In particular, we use penalties on truncated quadratic-polynomial bases to construct  $f_{drug}(x_{ij})$  and  $g_i(x_{ij})$ .

### Population or overall drug effect:

To formulate the PS-POP model we consider  $f_{drug}(x_{ij})$  as a smooth function which reflects the overall trend of dose and responses for the overall drug effect and is estimated by a penalized spline with quadratic truncated polynomial basis. Let  $k_1, k_2, \dots, k_K$  be a set of distinct knots in the range of  $x_{ij}$  and  $x_+ = \text{maximum}(0, x)$ . The number of knots  $K$  is fixed and large enough to ensure the flexibility of the curve. The knots are chosen as quantile of  $x$  with probability  $1/(K+1), \dots, K/(K+1)$ . Therefore, the functional presentation of  $f_{drug}(x_{ij})$  can be written as:

$$f_{drug}(x_{ij}) = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \sum_{k=1}^K u_k (x_{ij} - k_k)_+^2 \quad (3.3a)$$

### Subject-specific Biological replicate effect

A more flexible and more adaptable approach to model the subject-specific differences is the use of penalized regression semi-parametric technique. For a subject specific effect, the problem of smoothness is handled by applying truncated line bases (Durb`an, Harezlak, Wand, & Carroll, 2005). Here each subject-specific curve has a linear and a non-linear component to allow for more flexibility. We express  $g_i(x_{ij})$  in terms of truncated quadratic-polynomials assuming  $g_i$  as a smooth function estimated by penalized splines and specify number of knots. For simplicity we use same number of knots= $K$  as in the overall drug-effect curve and the formulation is as follows:

$$g_i(x_{ij}) = a_{i1} + a_{i2} x_{ij} + a_{i3} x_{ij}^2 + \sum_{k=1}^K v_{ik} (x_{ij} - k_k)_+^2 \quad (3.3b)$$

Where  $(a_{i1}, a_{i2}, a_{i3}) \sim N(0, \Sigma)$  and  $\Sigma$  is an unstructured  $3 \times 3$  matrix. Here the linear part is  $a_{i1} + a_{i2} x_{ij} + a_{i3} x_{ij}^2$  and the non-linear part is;  $\sum_{k=1}^K v_{ik} (x_{ij} - k_k)_+^2$ . Finally the PS-POP model is presented as in equation (3.4) below:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \sum_{k=1}^K u_k (x_{ij} - k_k)_+^2 + a_{i1} + a_{i2} x_{ij} + a_{i3} x_{ij}^2 + \sum_{k=1}^K v_{ik} (x_{ij} - k_k)_+^2 + \varepsilon_{ij} \quad (3.4)$$

where  $(a_{i1}, a_{i2}, a_{i3}) \sim N(0, \Sigma)$ ;  $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

In penalized spline regression model specification the smoothness of the population and subject-specific effect usually obtained by applying ridge penalty on  $u_k$  and  $v_{ik}$  to constrain their influence. The penalized spline smoother corresponds to the optimal predictor in a mixed model framework assuming  $u_k \sim N(0, \sigma_u^2)$  and  $v_{ik} \sim N(0, \sigma_v^2)$ . So, both components are considered as random and can easily be described in the mixed model framework. Brumback et.al (1999), Currie and Durban (2002), and Wand (2003) among others discussed the mixed model representation of penalized splines.

Our next step is to describe the model in the mixed model formulation according to Ruppert, Wand, & Carroll (2003) and Durban, Harezlak, Wand, & Carroll (2005). As described in Ruppert, Wand, & Carroll (2003), the connection between the penalized spline smoother and mixed model allows flexible modeling with correlated repeated data with smooth curves. In addition, with truncated polynomial bases, a mixed model representation is almost immediate from the form of the bases and the penalty function. The model 3.4 can be presented in the mixed model framework by the following matrix notations:

$$Y = X\beta + Zu + \varepsilon$$

where,

$$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{nn_i} \end{bmatrix}, X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, X_i = \begin{bmatrix} 1 & x_{i1} & x_{i1}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{in_i} & x_{in_i}^2 \end{bmatrix}$$

$$Z = \begin{bmatrix} Z_1 & X_1 & 0 & \cdots & 0 & Z_1 & 0 & \cdots & 0 \\ Z_2 & 0 & X_2 & \cdots & 0 & 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_n & 0 & 0 & \cdots & X_n & 0 & 0 & \cdots & Z_n \end{bmatrix}, Z_i = \begin{bmatrix} (x_{i1} - k_1)_+^2 & \cdots & (x_{i1} - k_K)_+^2 \\ \vdots & \ddots & \vdots \\ (x_{in_i} - k_1)_+^2 & \cdots & (x_{in_i} - k_K)_+^2 \end{bmatrix}$$

$$\beta = (\beta_0, \beta_1, \beta_2)^T, \quad u = [u_1, \dots, u_k, a_{11}, a_{12}, a_{13}, \dots, a_{n1}, a_{n2}, a_{n3}, v_{11}, \dots, v_{nK}]^T$$

$$G = Cov(u) = \begin{bmatrix} \sigma_u^2 I & 0 & 0 \\ 0 & \text{blockdianal } \Sigma_{1 \leq i \leq n} & 0 \\ 0 & 0 & \sigma_v^2 I \end{bmatrix}$$

The existing software for mixed model analyses makes it possible to implement the complicated penalized spline regression models in a simple mixed model representation and allows us to fit the above semi-parametric mixed model using the R-package ‘nlme’ (Pinheiro, Bates, DebRoy, & Sarkar, 2009).

A standard estimation criterion for variance component Restricted Maximum Likelihood (REML) is used to estimate model parameters. For model in 3.5

$$l_R(\sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2) = \frac{1}{2} \log|V| - \frac{1}{2} \log|X^T V^{-1} X| - \frac{1}{2} y^T (V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}) y$$

where  $V = ZGZ^T + \sigma_\varepsilon^2 I$  and  $G$  is defined above. The vector of parameters  $\beta$  and the random coefficients vector  $u$  can be determined using the prediction:

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

$$\hat{u} = \hat{G} Z^T \hat{V}^{-1} (y - X \hat{\beta})$$

According to Ruppert, Wand, & Carroll (2003) variance calculation should be done with respect to the conditional distribution  $y|u$ . Originally, Ruppert, Wand, & Carroll

(2003) derived the standard deviation of  $\{\hat{f}(x)|u\}$ . We follow the same formulation for our PS-POP model as below:

$$\widehat{SD}\{\hat{f}(x)|u\} = \widehat{\sigma}_\epsilon \sqrt{C_x(C^T C + \frac{\widehat{\sigma}_\epsilon^2}{\widehat{\sigma}_u^2} D)^{-1} C^T C (C^T C + \frac{\widehat{\sigma}_\epsilon^2}{\widehat{\sigma}_u^2} D)^{-1} C_x^T}$$

where  $C_x = [X_x \ Z_x]$ ,  $D$  is some symmetric positive semi definite matrix associated with the penalty, and  $C = [C_{xi}]_{1 \leq i \leq n}$ , and the corresponding  $100(1-\alpha)\%$  confidence limit (CL) is as follows:

$$\hat{f}(x) \pm z \left(1 - \frac{\alpha}{2}\right) \widehat{SD}\{\hat{f}(x)|u\}$$

### 3.3.2 PSDR-PS- model for individual subject curves:

To formulate the PS model for individual subject-specific curves, we considered standard penalized model for individual  $i^{\text{th}}$  subject. The knots and penalty are defined as in Section 3.3.1. Since we are considering truncated lines as the basis for our penalized spline regression the mathematical form of the PS model can be represented for an individual  $i^{\text{th}}$  subject:

$$Y_{ij} = f(x_{ij}) + \epsilon_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \sum_{k=1}^K u_k (x_{ij} - k_k)_+^2 + \epsilon_{ij} \quad (3.5)$$

where  $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ . Here equation (3.5) is a smooth function which is estimated by a penalized spline regression model. This model assumes the individual subject curve has two components. One is linear (here, we are considering quadratic form),  $\beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2$  and the other part is non-linear presented as  $\sum_{k=1}^K u_k (x_{ij} - k_k)_+^2$ . For this simple penalized spline regression model, we will briefly show the penalized spline regression fit and then the representation of the mixed model framework. Here  $f(x)$  is our

smooth function and we apply a quadratic penalty on the basis coefficients to control model smoothness and  $\lambda$  is the associated smoothing parameter. Then the roughness penalty term is,  $\lambda\beta^T D \beta$ , where D is a  $(K + 3) \times (K + 3)$  matrix. The estimates of  $\beta = (\beta_0, \beta_1, \beta_2)$  and  $u = (u_1, \dots, u_K)$  are obtained by minimizing the penalized least squares which can be written as:

$$\sum_{j=1}^{n_i} \{Y_{ij} - (\beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \sum_{k=1}^K u_k (x_{ij} - k_k)_+^2)\}^2 + \lambda\beta^T D \beta$$

The minimization criteria can easily be described in the mixed model representation where penalized spline smoother correspond to the optimal predictor in a mixed model framework assuming  $u_k \sim N(0, \sigma_u^2)$  and  $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_u^2}$ .

Let

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{and} \quad u = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}$$

be the coefficients of the polynomial functions and truncated line functions, respectively.

The corresponding vectors can be defined for  $i^{\text{th}}$  subject:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{n_i} & x_{n_i}^2 \end{bmatrix}; \quad Z = \begin{bmatrix} (x_1 - k_1)_+^2 & \cdots & (x_1 - k_K)_+^2 \\ \vdots & \ddots & \vdots \\ (x_{n_i} - k_1)_+^2 & \cdots & (x_{n_i} - k_K)_+^2 \end{bmatrix}$$

The penalized spline fitting criteria become,

$$\|y - X\beta - Zu\|^2 + \lambda \|u\|^2$$

Treating  $u$  as a set of random coefficients with  $cov(u) = \sigma_u^2 I$  where  $\sigma_u^2 = \sigma_\varepsilon^2 / \lambda$

The mixed model representation is:

$$y = X\beta + Zu + \varepsilon; \quad \text{cov} \begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sigma_u^2 I & 0 \\ 0 & \sigma_\varepsilon^2 I \end{bmatrix}$$

Fitted values:  $\hat{f}(x) = C(C^T C + \lambda D)^{-1} C^T y$

where  $C = [X \ Z]$ ;  $D = \text{diag}(0,0,0,1,1, \dots, 1)$ ;  $\lambda = \sigma_\varepsilon^2 / \sigma_u^2$ .

The standard deviation of  $\{ \hat{f}(x) | u \}$  for individual  $i^{\text{th}}$  subject is

$$\widehat{SD}\{\hat{f}(x)|u\} = \widehat{\sigma}_\varepsilon \sqrt{C_x (C^T C + \frac{\widehat{\sigma}_\varepsilon^2}{\widehat{\sigma}_u^2} D)^{-1} C^T C (C^T C + \frac{\widehat{\sigma}_\varepsilon^2}{\widehat{\sigma}_u^2} D)^{-1} C_x^T}$$

with  $C_x = [X_x \ Z_x]$  and  $C = [C_{xi}]_{1 \leq i \leq n}$ , and the corresponding 100(1- $\alpha$ )% confidence limit (CL) is as follows:

$$\hat{f}(x) \pm z \left(1 - \frac{\alpha}{2}\right) \widehat{SD}\{\hat{f}(x)|u\}$$

### 3.3.3 Algorithm to calculate relative IC<sub>50</sub> from the PS-POP and PS model

The next step in our PSDR dose-response analysis is to obtain the quantities of interest, half-maximal inhibitory concentration (IC<sub>50</sub>) and their associated properties. Commonly used models in dose-response analysis are those generating from sigmoidal functions (for example, LL4, EXD3 (Table 3-2)), and one clear advantage of the LL4 and EXD3 model is that one of the model parameter is defined as the dose which gives the percent of response( for example, IC<sub>50</sub> ). Once the dose-response curve is fitted with the sigmoidal model, the IC<sub>50</sub> and the SD of the IC<sub>50</sub> is obtained from the estimated parameter of interest.

For PSDR models, we estimated dose response curves with CL using mixed model formulation. The respective summary measures of dose response curves, such as IC<sub>50</sub> and SD of IC<sub>50</sub>, needs to be numerically estimated from the fitted response. We present an



algorithm to estimate relative  $IC_{50}$  and SD of  $IC_{50}$  from the fitted PSDR model. Originally, the algorithm was developed to estimate relative  $IC_{50}$  which is the concentration required bringing the curve down to point half way between the top and bottom plateaus of the curve. This algorithm can be used for any quantile of interest (for example, 90<sup>th</sup> percentile).

The relative  $IC_{50}$  algorithm starts with estimating the fitted responses from the PSDR model and involves the following steps:

- 1) Fit the PSDR model using mixed model representation, and obtain the fitted responses.
- 2) Estimate the mid value from the fitted responses: mid-value= (maximum of fitted-value – minimum of fitted-value)/2.
- 3) Specify the function for which a root is needed. Here, the function is the fitted PSDR minus the mid-value which is estimated in Step 2.
- 4) Specify the lower and upper end points of the interval to search for a root (i.e., zero) of the function in Step 3. The upper endpoint must be strictly larger than the lower endpoint and the function values at the endpoints must be of opposite signs (or zero).
- 5) The root of the equation is searched by the method based on the algorithms given by Brent (1973). The algorithm assumes a continuous function (which then is known to have at least one root in the interval).
- 6) Convergence is declared either if  $f(x) = 0$  or the change in  $x$  for one step of the algorithm is less than tolerance.

The estimated  $IC_{50}$  is obtained from the above algorithm and we calculate the SD of estimated  $IC_{50}$  utilizing the "Delta Method" (Oehlert, 1992) in Equation (3.6).

$$Var(IC_{50}) = Var(\hat{f}(IC_{50}) \times \{\hat{f}'(IC_{50})\}^{-2}) \quad (3.6)$$

Where  $\hat{f}'(IC_{50})$  is the 1<sup>st</sup> derivative of the estimated responses at  $x = IC_{50}$ .

### 3.4 Simulation studies

The key objective of the simulation study was to fit dose-response curves, estimate  $IC_{50}$  and compare four (LL4, EXD3, QUAD and PS) candidate parametric models with the PS-POP model. Two hundred samples, each with fifty subjects, were generated and evaluated to 12 dose levels. The parameters and 12-point dose levels used in simulation studies were designed to mimic the observed values from the malaria drug study (Section 3.5). The proposed PSDR models (PS and PS-POP), described in Section 3.3, were fitted for each simulation study, and relative  $IC_{50}$  were estimated using the algorithm in Section 3.3.3.

The performance of the PS-POP model was investigated with regards to:

- 1) fitted-MSE (fitted-mean squared error = {fitted response-true response}<sup>2</sup>) from the candidate models and
- 2)  $IC_{50}$ -MSE ( $IC_{50}$ -mean squared error = {estimated  $IC_{50}$ -true  $IC_{50}$ }<sup>2</sup>).

The MSE were estimated for each simulated data and then averaged across the 200 simulated data to obtain fitted-MSE and, similarly,  $IC_{50}$ -MSE. In the simulation studies, we first explore generating data with a single dose-response function, LL4, presented here as “Sim-LL4”. Next, we explored the data generated with a mixture of functions which could arise from a subject that carries two populations of infectious agents with different levels of sensitivity to the drug. For simplicity, we simulated mixtures of two functional

forms. Let  $f_1(x)$  denote the density associated with the function  $D_1$  and  $f_2(x)$  denote the density associated with the function  $D_2$ .

The overall mixture density function is given by

$$f_{Mix}(x) = p f_1(x) + (1 - p) f_2(x)$$

where,  $p$  (mixing percentages) is the contamination percentage (expressed as a fraction, so  $p = 0.50$  corresponds to 50% /50% mixture). We chose to generate  $f_i(x)$   $i = 1,2$  with the functions presented in Table 3-2 with fifty-percent QUAD and fifty-percent EXD3 functions (“Mix-Quad-EXD3”). Additionally, we chose to generate data with 50% samples from LL4 function and 50% samples from the EXD3 (“Fifty/Fifty-LL4-EXD3”). Figure 3-2 represents a boxplot summary of the results from simulated data generated through three different scenarios: Sim-LL4, Mix-QUAD-EXD3, and Fifty/Fifty-LL4-EXD3.

Using the above simulated data, the proposed PS-POP model was compared with four candidate models: LL4, EXD3, QUAD and PS. Curves fitted through the PS-POP model demonstrated reduced fitted-MSE among the five candidate models (Figure 3-2: a1, b1, c1) and reduced  $IC_{50}$ -MSE (Figure 3-2: a2, b2, c2). For example, the simulation study (Figure 3-2: b2) generated from “Mix-Quad-EXD3” showed that the ratio of the  $IC_{50}$ -MSE of the proposed PS-POP model over LL4 is 0.68 (similarly, EXD3= 0.53; QUAD= 0.12; PS= 0.61). Therefore, the reduction of the  $IC_{50}$ -MSE, measured by  $(1-\text{ratio}) \times 100$ , by the proposed PS-POP model over LL4, EXD3, QUAD, and PS are 68%, 56%, 83% and 61%, respectively. From the same simulation study (Figure 3-2: b1), the reduction of Fitted-MSE by PS-POP model over LL4 is 32% (EXD3=46%, QUAD=88% and PS=39%), which suggests efficiency gain of estimating  $IC_{50}$  as well as the fitted curve by properly

accounting for the correlation and overall population effect by the proposed PS-POP model.

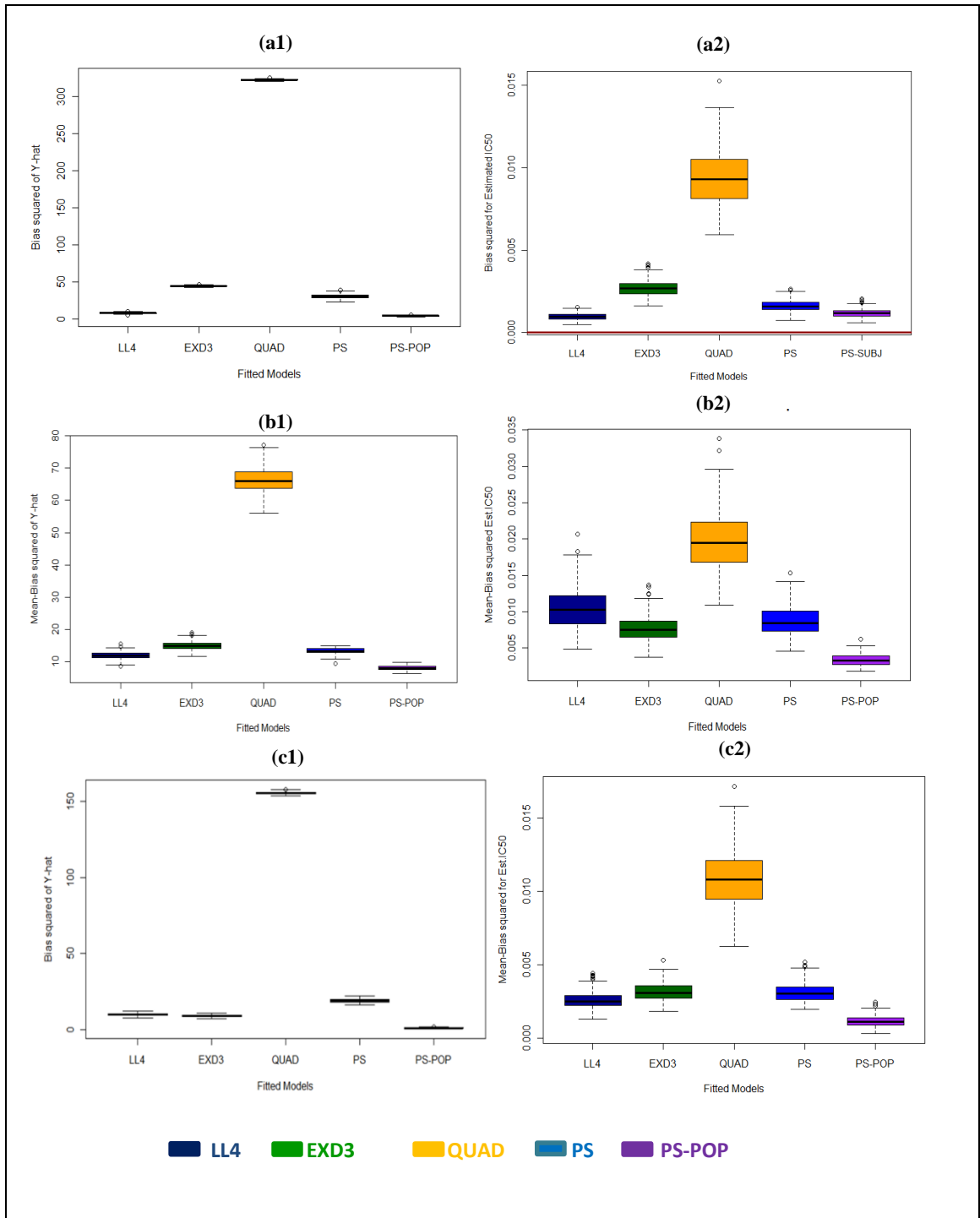


Figure 3-2 Boxplot summaries of the three simulation studies. Left-panels represent ‘fitted-MSE’ from all five candidate models from the three simulation studies. Right-panels represent ‘IC<sub>50</sub>-MSE’ from three simulation studies. Top-row: Box plot summaries from the simulation study “Sim-LL4”, where (a1) represents fitted-MSE and (a2) represents IC<sub>50</sub>-MSE. Similarly, second-row (b1-b2) study “Mix-QUAD-EXD3” and Bottom-row (c1-c2) study “Fifty/Fifty-LL4-EXD3”. Five candidate models: LL4, EXD3, QUAD, PS and PS-POP.

Additionally, the PS-POP model gained efficiency over the four candidate models in estimating  $IC_{50}$  when data was generated from “Fifty/Fifty-LL4-EXD3” (Figure 3-2: c2) and also gained efficiency over the EXD3, QUAD and PS models when data was generated from “Sim-LL4” (Figure 3-2: a2). Analysis through the well-known dose-response model-LL4 fitted the data well but could not achieve efficiency in estimating  $IC_{50}$ , especially when the source of the data did not follow a sigmoidal shape (Figure 3-2: b2, c2).

Similarly, the PS model which accounted for within-subject correlation, also resulted in smaller  $IC_{50}$ -MSE over LL4 and QUAD in estimating  $IC_{50}$ .

Based on the simulation study results, the PS-POP model offers an improved, flexible method of dose-response analysis which uses a data-driven approach, accounts for overall population effect, and considers subject-specific correlation.

### 3.5 PSDR model: application to malaria data

We re-analyzed data from the malaria *ex vivo* drug sensitivity assay used for surveillance of resistance to drugs AMQ, CQ and MEF as described in Section 3.2.

A total of 106 patients enrolled in this study with a median age of 23 (IQR: 18- 28). For each individual drug, dose-response curves were fitted with confidence intervals, and the corresponding  $IC_{50}$  with SD were estimated using the proposed PS-POP model. In addition, LL4, EXD3, QUAD, and PS models were fitted for 106 individual subjects and related  $IC_{50}$  were estimated for each drug separately. Figure 3-3 (a-d) represents results from malaria data analysis. When analyzing AMQ data with the PS-POP model, the estimated median  $IC_{50}$  is 1.06 nM with IQR: 0.95-1.15. The variability of estimated  $IC_{50}$  obtained from the proposed PS-POP model was the smallest among the four fitted models: LL4, EXD3, QUAD, and PS. In addition, the proposed PS model also performed better in

estimating  $IC_{50}$  (median 1.0 nM with IQR: 0.74-1.19) with less variability compared to LL4, EXD3 and QUAD, model (Figure 3-3 a). The estimated dose-response curves showed a strong uphill linear pattern between the five candidate models (Figure 3-3 b). Furthermore, the PS and PS-POP model appropriately fitted the dose-response curve for the second subject discussed in the introduction section (Figure 3-3 d) and precisely estimated  $IC_{50}$  (PS:  $1.0 \pm 0.14$  nM; PS-POP:  $1.26 \pm 0.04$  nM) with a small SD.

The standard sigmoidal model fits the dose-response curves well when the pre-specified assumptions of the functions are fulfilled. The key conditions are that the sigmoidal function is monotonic, symmetric about the  $IC_{50}$ , and the dose and  $IC_{50}$  are easily estimable from the model parameter. The pattern of the observed dose-response relationship of the malaria data we examined in our example did not fulfill the assumptions of the sigmoidal functions. For example, some samples were asymmetric or non-monotonic. As a consequence, our preliminary analysis with sigmoidal models like LL4 yielded either a poor fitting or unreliable estimated  $IC_{50}$ . Thus, the appropriate estimation of  $IC_{50}$  and unbiased fitting of malaria ex vivo drug sensitivity assay with PS-POP models performed efficiently when the observed pattern of the dose-response curve was asymmetric or non-monotonic. Moreover, when the observed data were symmetric and monotonic, PS-POP models yielded curve fitting and estimated  $IC_{50}$  comparable to parametric sigmoidal models.

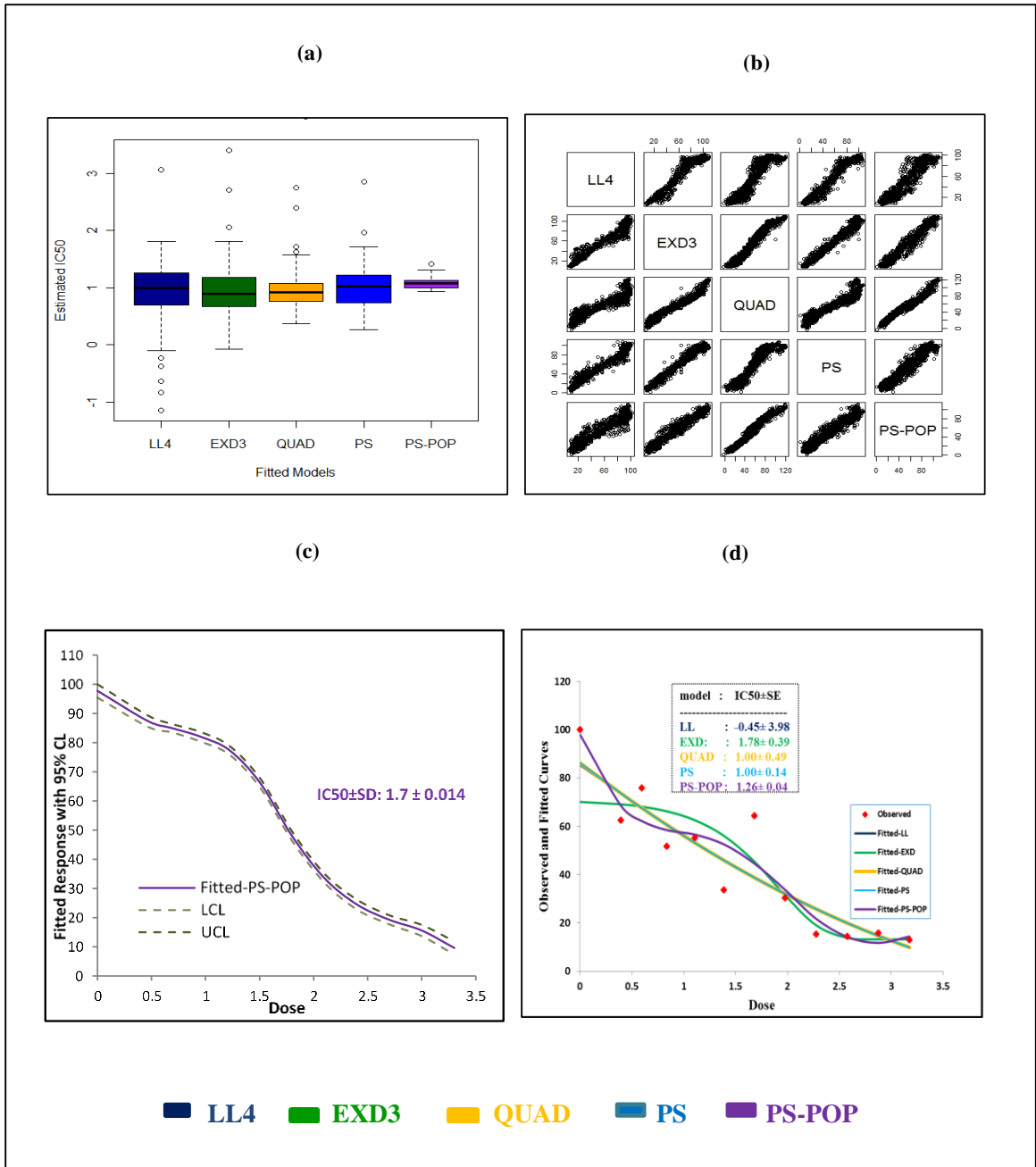


Figure 3-3 Summary results from the data in AMQ (a) Boxplot Summary of estimated  $IC_{50}$  by five candidate models (b) Scatter plot of correlation of the fitted curves among the five candidate models (c) Plot of an example fitted curve with 95% CL obtained from the PS-POP model and (d) Plot of an example individual: observed data and fitted curves from five candidate models. Five candidate models: LL4, EXD3, QUAD, PS and PS-POP.

It is evident that the robust PSDR modeling strategy appropriately estimated the observed pattern of the real dose-response relationship whether or not sigmoidal model



fulfilled the underlying conditions or even the dose-response function didn't work due to the restriction of the model assumptions. In addition, PS-POP models estimated  $IC_{50}$  with the smallest variability compared to the candidate models.

### 3.6 Conclusion and Discussion

Estimating dose-response curves or  $IC_{50}$  with sigmoidal models may produce inaccurate results if the data does not follow a specific parametric shape. Those inaccurate results may be significant enough to result in incorrect decisions on the target dose of a drug. The PSDR models show promise as a method that can be used alongside the parametric analysis of dose-response data and as a tool for curve fitting and effective dose estimation when the sigmoidal model is inadequate. Based on the results from the simulation studies, the PSDR model offers an improved, flexible method of dose-response analysis which uses a data-driven procedure and considers subject-specific correlation.

In practice, if the observed data is asymmetric or non-monotonic or does not meet the model assumptions, then those data are deleted in order to generate the fitted curves. In this process, some important information might be lost by deleting these atypical observations. Since PSDR models use all data points as much as possible, it reduces the uncertainties and identifies the areas where data gaps exist.

In summary, we conclude that PSDR method provides a robust modeling approach to that of LL4, EXD3, and QUAD. Moreover, the estimation of  $IC_{50}$  is precise and is able to produce expected estimates with minimum SD. The PSDR model shows an advantage over the well-liked LL4 or EXD3 models in several ways:

- 1) Overall drug effect can be estimated in addition to the subject-specific estimate
- 2) Fits monotonic data as well as asymmetric and non-monotonic data

- 3) Important dose-response features will not be omitted
- 4) Adequately estimates effective dose ( i.e.  $IC_{50}$  with minimum SD).

Since the results we obtained are based on a large sample theory, a potential limitation of our proposed PSDR model is that a reasonable sample size may be needed. Also, in our research we did not apply testing the monotonicity of the dose-response curves which are of practical interest when the observed samples are non-monotonic. Our future goal is to develop a monotonicity test before applying the PSDR model.

The PSDR method can also be used for other dose-response modeling scenarios. Our semi-parametric model is better suited than traditional sigmoidal models to fit other nonparametric curves such as J-shaped and U-shaped curves frequently observed in toxicology and epidemiology studies.

Our approach for estimating  $IC_{50}$  can also be used to estimate the half-maximal  $EC_{50}$  (effective concentration) and the  $LD_{50}$  (lethal dose 50%) or the  $LC_{50}$  (lethal concentration) and time of a toxic substance or dose needed for radiation to kill half the tested population.

## REFERENCES

- Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.
- Bretz, F., Pinheiro, J. C., & Branson, M. (2005). Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. *Biometrics*, 61 , 738-748.
- Brumback, B., Ruppert, D., & Wand, M. (1999). Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior: Comment. *Journal of the American Statistical Association*, 94:794–797.
- Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 829-836.
- Currie, I. D., & Durb`an, M. (2002). Flexible smoothing with P-splines: A unified approach. *Statistical Modelling*, 2:333–349.
- Desquilbet, L., & Mariotti, F. (2010). Dose-response analyses using restricted cubic spline functions in public health research. *Statistics in Medicine*, 1037–1057.
- Djeundje, V. A. (2010). Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, 1202-1224.
- Durb`an, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24:1153–1167.
- Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–102.
- Fan, J., & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, 2008-2036.

- Govindarajulu, U., Malloy, E., Ganguli, B., Spiegelman, D., & Eisen, E. (2009). The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *Int J Biostat*, 5(1):Article 2.
- Greenland, S. (1995). Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, 6(4):356-65.
- Ndiaye, D., Patel, V., Demas, A., LeRoux, M., Ndir, O., Mboup, S., & Wirth, D. (2010). A non-radioactive DAPI-based high-throughput in vitro assay to assess Plasmodium falciparum responsiveness to antimalarials--increased sensitivity of P. falciparum to chloroquine in Senegal". *AJTMH*, 82(2): 228-230.
- Ngo, L., & Wand, M. (2004). Smoothing with Mixed Model Software. *Journal of Statistical Software*, 9(1-54), 1-54.
- Oehlert, G. W. (1992). A Note on the Delta Method. *The American Statistician*, Vol. 46, No. 1, p. 27-29.
- Pinheiro, J. C., Bates, D. M., DebRoy, S., & Sarkar, D. a. (2009). *nlme: Linear and nonlinear Mixed Effects Models*. version 3.1-96: R package.
- R Development Core Team. (2008). *R: A language and environment for statistical computing*. Retrieved from R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0: <http://www.R-project.org>.
- Ruppert, D., Wand, M. P., & Carroll, R. (2003). *Semiparametric Regression*. New York:: Cambridge University Press.
- Steenland, K., & Deddens, J. (2004). A practical guide to dose-response analyses and risk assessment in occupational epidemiology. *Epidemiology*, 15:63-70.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18:223–249.

## CHAPTER 4. PS-SIZER: A VISUAL TOOL TO INVESTIGATE SIGNIFICANT FEATURES IN LONGITUDINAL DATA

### Summary

We propose Penalized Spline Significant Zero Crossings of Derivatives (PS-SiZer), an extension of SiZer (Chaudhuri & Marron, 1999) as a graphical tool for the exploratory analysis of the longitudinal data. The standard implementation of SiZer is based on the local linear smoother with a kernel-type smoothing method for curve estimation problems. In longitudinal studies data are often correlated, and it is necessary to account for the within-subject correlation. In our research, we propose an extension of the SiZer methodology for correlated observations arising from longitudinal settings by using a computationally efficient smoothing method, a penalized spline regression model. We apply our PS-SiZer methodology to analyze differential pattern of body weight change over time among HIV patients using data from the International Epidemiologic Database to Evaluate AIDS (IeDEA) collaboration.

### 4.1 Introduction

In various clinical and epidemiological studies, longitudinal data are frequently collected over time for several subjects. Such data are often large in size and are subject to within-subject correlation among repeated measurements from the same subject over time. An important first step before performing any kind of statistical analysis is to familiarize oneself with the data at hand (this is often called exploratory data analysis). This usually involves graphing the variables in various distributional displays. The most common way of visualizing longitudinal data is via cross-sectional summaries or spaghetti plots.

While working on the large HIV data arising from IeDEA collaboration, the exploratory data analysis could be useful to explore the data for comprehensive visualization so that we could understand the pattern of weight change among patients receiving antiretroviral therapy (ART). The research data consists of more than 185,000 HIV patients with more than 2 million observations collected over a 4-year follow-up period. Commonly available data visualization graphical tools were used to investigate weight trajectories over time. For example, only 1% of HIV data was extracted to generate a sample scatter plot, spaghetti plot, and observed median plot over time (Figure 4-1 a-c). The graphs depicted observed weight measurements against weeks since the start of ART. Although these tools are useful for fewer subjects, the weight trajectory or features of the data become obscured for large number of patients and/or measurements collected over a 4-year follow-up period. The scatter plot and spaghetti plot become too busy, the median plot did not reveal any specific pattern of weight change over time. In absence of proper understanding of the pattern of the response over time via exploratory visualization, modeling of the mean response curve has become a challenge.

In recent years, modeling longitudinal trajectories utilizing smoothing methods has gained a lot of attention. Smoothing methods are known for their flexibility and can be used to detect the pattern of change in responses over time (Figure 4-1 d). A large number of research studies describing various approaches of smoothing techniques is available in recent literature. A few examples include Green & Silverman (1994), Wahba (1991), Brumback & Rice (1998), Eubank (2000), Ruppert, Wand & Carroll (2003) and the references therein.

As noted in Marron (1996), a hurdle in the application of the smoothing method is the selection of the smoothing parameter. The question is: “what is the ‘best’ estimate of the smoothing parameter to reveal the true structure or feature of the underlying curves?” Marron (1996) mentioned that the statistical inference is challenging at a single smoothing level because interesting features that are present in the data may appear at some levels of smoothing, whereas some features may disappear by over-smoothing or under-smoothing. Especially when data is large, finding an estimate of optimum or best smoothing parameter become more difficult. SiZer map (Chaudhuri & Marron, 1999) was proposed to address these issues as an exploratory data analysis tool to reveal various features in the data at various levels of smoothing. SiZer map has gained popularity by different extensions such as Robust SiZer (Hannig & Lee, 2006), Quantile SiZer (Park, Lee, & Hannig, 2010), and various SiZer in time series data (Hannig, Lee, & Park, 2012).

The application of SiZer map is not common in biomedical research, especially in longitudinal settings. In our research, we advocate for the use of SiZer map as an alternative and complementary visual exploratory tool for large longitudinal data settings. We propose an extension of SiZer methodology to account for the within-subject variability that arises in repeated measurements. The SiZer map combines statistical inferences to reveal which features are really present with a color-coded graphical map that makes the tool more appealing and makes the tool quickly comprehensible and accessible.

The Significant Zero Crossings of Derivatives (SiZer) (Chaudhuri & Marron, 1999) is a useful visualization tool for understanding the significant features of smoothing curves. The standard implementation of SiZer (Chaudhuri & Marron, 1999) is based on the local linear smoother with a kernel-type smoothing method for bivariate data. SiZer



simultaneously studies a family of smooth curves under a wide range of smoothing parameters (bandwidths in kernel smoothing setting), and the inference is focused on a smoothed version of the underlying curve viewed at varying levels of bandwidths. The statistical inference is based on the derivatives (slope) of the smooth curve by constructing a confidence limit (CL) at each location and also at each level of the smoothing parameter. The SiZer map represent the structures of the curves at various level of smoothing by using a two-dimensional plot where the horizontal axis represents the location (i.e., time), while the vertical axis represents the scale (various level of smooth). The SiZer map classifies every point along the horizontal axis into one of the three states: the estimated slope is positive (i.e., the CL of the first derivative contains only positive values), negative (the CL contains only negative values), or possible zero (the CL contains zero).

A number of advancements have already taken place in the development of SiZer. Hannig & Marron (2006) put forward an advanced distribution theory for SiZer to make inference better by substituting the appropriate quantile for the confidence interval to account for the multiplicity issue. Another important SiZer tool was proposed by Park & Kang (2008) and is capable of comparing numerous curves on the basis of their dissimilarities of smoothness in independently observed data. Park, Marron, & Rondonotti (2004) recommended a dependent SiZer. This particular dependent SiZer extends the methodology into the area of the time series and employs an implicit auto-covariance function when applying goodness of fit tests. Rondonotti, Marron, & Park (2007) extended the SiZer to time series and formed a technique that is more supple and capable in accounting for the dependence structure existing in the data so as to notice considerable features by not presuming, but instead approximating, the auto-covariance function.

Recently, a more improved version of SiZer was developed by Park, Hannig, & Kang (2009) for time series by adding the extreme value hypothesis which was put forward by Hannig & Marron (2006). The method proposed in the paper (Park, Hannig, & Kang, 2009) was to obtain a quantile that decreases the number of unwanted false pixels. They also suggested a new auto-covariance estimator by means of a varied time series. This improved model also does not depend on pilot residuals and bandwidths from an approximation like in Rondonotti, Marron, & Park (2007).

The existing SiZer based methods including the time series extensions do not account for the subject-specific variability arising from repeated measurements in a longitudinal setting. Therefore, application of standard SiZer and its extensions in longitudinal studies may be limited or misleading.

Lately semi-parametric approaches have emerged as a flexible means to model longitudinal data. The time course is often too complex to model parametrically; hence, in recent years, semi-parametric analysis of longitudinal data has gained traction. The parametric model may have limitations in detecting fine features, whereas semi-parametric regression, such as penalized spline regression applied to longitudinal data allows for exploration of the unknown shape of the mean curve and detects subject-specific deviation from the mean curve (Staniswalis & Lee, 1998).

Penalized spline regression possess computational advantages over non-parametric kernel smoothing or regression spline or smoothing spline methods. For large data, penalized spline regression is computationally less expensive than smoothing spline and kernel based method. In addition the mixed model representation of penalized spline allows for a seamless fusion between parametric mixed model and smoothing. Such models are

flexible to incorporate within subject correlation arises from repeated measure as a random component. Therefore, penalized spline regression models are good candidates for incorporating in the SiZer map to account for subject specific correlation.

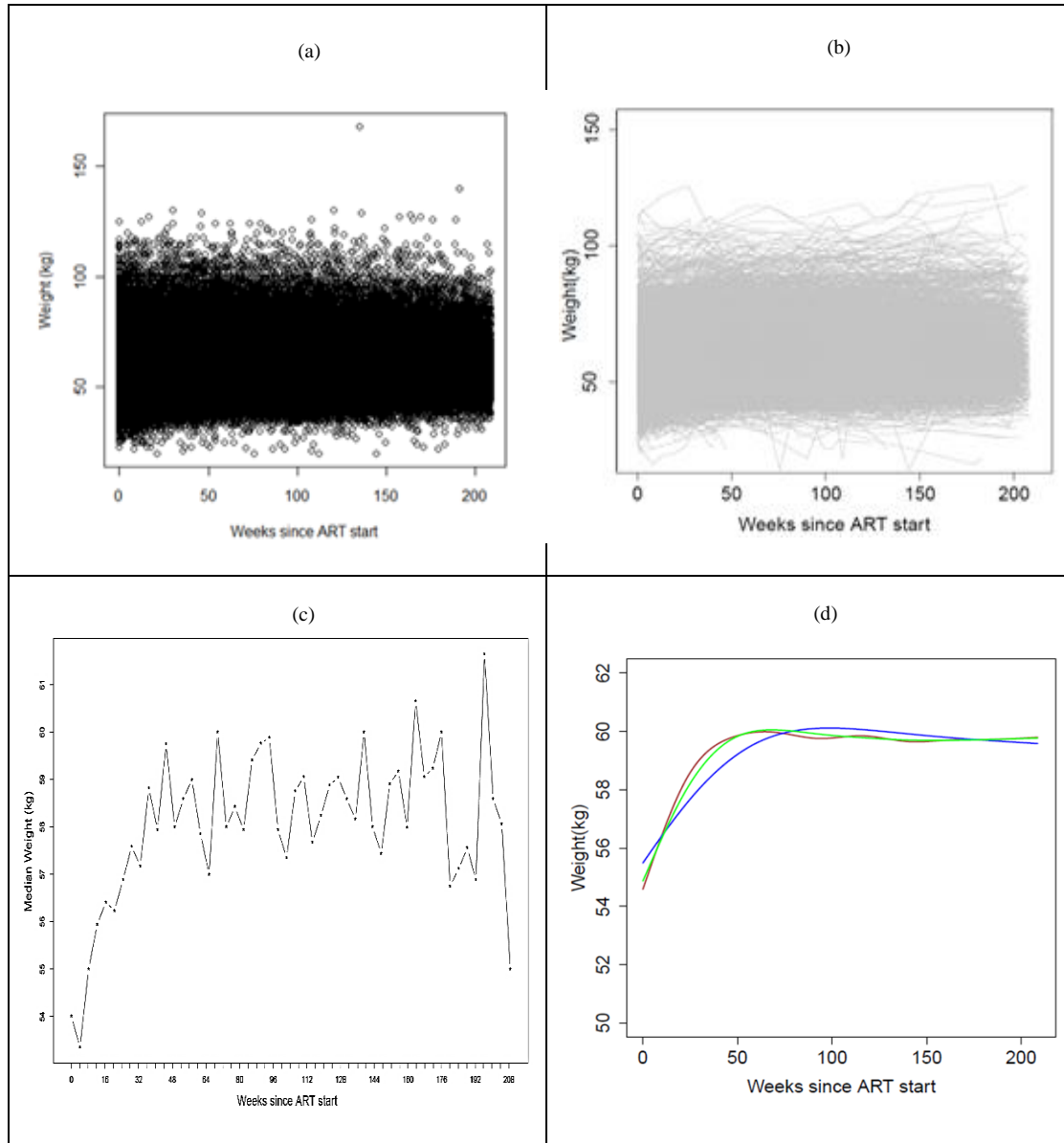


Figure 4-1 Upper-left panel represents scatter plot; Upper-right panel represents spaghetti plot; Lower-left panel represent median weight (kg) over time; Lower-right panel represents smoothing curve at 3 different smoothing level. Data is from 1% of HIV patients from IeDEA (2,000 patients, 12,000 observations).

The motivation of this research work is to analyze HIV data arising from the IeDEA study (details in Section 4.4). The body weight of HIV patients was collected as a repeated measurement for each patient in IeDEA. The clinical interest is to describe the pattern of body weight changes among patients receiving ART and to assess the impact of the ART regimens as a surrogate for the treatment effectiveness in HIV patients. Since weight measurements are the longitudinally collected markers, it is important to opt for a pertinent weight trajectory. The observed HIV weight change data is curvilinear; therefore, we want to detect the fine features (i.e. bumps and valleys) in the function in order to explore the features of the change of weight over time. In order to explore the HIV data precisely, our research aims to extend the SiZer tool allowing for subject-specific correlation in the model. The proposed extension, named Penalized Spline SiZer (PS-SiZer), utilizes the penalized spline regression method. The proposed approach achieves the following: (1) extends SiZer to investigate significant features in mean regression function arising from longitudinal data accounting for correlation in the analysis and (2) enhances the underlying smoothing model used in standard SiZer by a computationally efficient smoothing model for correlated data.

The paper is organized as follows. In Section 4.2, we give a brief overview of the core ideas of the SiZer methodology proposed by Chaudhuri & Marron (1999) and the spline SiZer of Marron & Zhang (2005). In Section 4.3, we provide the development of the proposed PS-SiZer map procedure for longitudinal data. Simulation studies are presented in Section 4.4. The analysis of the IeDEA data is summarized in Section 4.5. We conclude in Section 4.6 with a brief discussion of our results.

## 4.2 SiZer method

In this section, we provide core ideas of the SiZer methodology originally proposed by Chaudhuri & Marron (1999). For a given set of observed data  $\{(x_i, y_i)_{i=1}^n\}$  and a smoothing function  $g(x)$ , we can consider a non-parametric regression model as below:

$$y_i = g(x_i) + \epsilon_i, \quad i = 1, \dots, n, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2),$$

where  $g(x)$  is some ‘smooth’ regression function that needs to be estimated from the set of observed data  $(x_i, y_i)$ , and  $\epsilon_i$  is the random error component with variance  $\sigma_\epsilon^2$ .

Using the above modeling framework, we briefly present two variants of SiZer: the standard SiZer based on local linear smoother by Chaudhuri & Marron (1999) and extended with advanced theory by Hannig & Marron (2006) and the spline SiZer by Marron & Zhang (2005). In this paper, the standard SiZer map (Chaudhuri & Marron, 1999) based on a local linear smoother is named as LL-SiZer, and the smoothing spline SiZer (Hannig & Marron, 2006) is referred to as SS-SiZer. Brief details are in the following subsections.

### 4.2.1 LL-SiZer

The smooth function,  $g(x)$ , is estimated using a non-parametric regression method with smoothing parameter (bandwidth)  $\lambda$  and it is denoted by  $\hat{g}_\lambda(x)$ . The LL-SiZer applies the local linear regression (Fan & Gijbels, 1996) to estimate  $g(x)$  and its derivative,  $g'(x)$ . Local linear smoother estimate of  $g(x)$  at a smoothing level  $\lambda$  at each location of  $x$   $\hat{g}_\lambda(x)$  given by

$$\hat{g}_\lambda(x) = \underset{a_0, a_1}{\operatorname{argmin}} \sum_{i=1}^n [y_i - \{a_0 + a_1(x_i - x)\}]^2 \times K_\lambda(x - x_i)$$

where  $\operatorname{argmin}$  means minimizing jointly over regression coefficients,  $a_0$  and  $a_1$  at each point “ $x$ ”. A line is fitted to the data for each  $x$  using  $K_\lambda$  weighted least-squares. Here  $K(\cdot)$

is a kernel taken as the standard normal density function and denoted by  $K_\lambda = K(\cdot/\lambda)/\lambda$ . Let  $g_\lambda(x) = K_\lambda * g(x)$  and  $g'_\lambda(x) = K'_\lambda * g(x)$ . Then the estimates of  $g_\lambda(x)$  and  $g'_\lambda(x)$  are given as  $\hat{g}_\lambda(x) = \hat{a}_0$  and  $\hat{g}'_\lambda(x) = \hat{a}_1$ , respectively. Therefore we can construct a family of smooth with estimated regression functions at different levels of  $\lambda$  values.

The LL-SiZer model specification (Chaudhuri & Marron, 1999) considered a family of smooths with the smoothing parameter  $\lambda: \{\hat{g}_\lambda(x): \lambda \in [\lambda_{min}, \lambda_{max}]\}$ . The LL-SiZer (Chaudhuri & Marron, 1999) considered  $\lambda_{min}$  to be the smallest bandwidth for which there is no substantial distortion in construction of the binned implementation of the smoothers, such that,  $\lambda_{min} = 2 * (binwidth)$  and  $\lambda_{max}$  is the range of the data.

A SiZer map is constructed by summarizing the results of a sequence of hypothesis tests for each pair of  $(x, \lambda)$ . The test is performed by constructing confidence limit (CL) for  $g'_\lambda(x)$ . The confidence limits are obtained as follows:

$$\hat{g}'_\lambda(x) \pm q_{1-\alpha}(\lambda) * \widehat{sd}(\hat{g}'_\lambda(x)) \}$$

where,  $\hat{g}'_\lambda(x) \equiv d[\hat{g}_\lambda(x)]/dx$ ,  $\widehat{sd}(\hat{g}'_\lambda(x))$  is the estimated standard error of  $\hat{g}'_\lambda$ , and  $q_{1-\alpha}(\lambda)$  is defined as an appropriate quantile,  $\alpha$  say 5% significance limit.

A SiZer map is a two dimensional plot, where the horizontal axis of the map represents the location  $x$  and the vertical axis represents the bandwidth  $\lambda$ . Then significance feature can be obtained from every point along the axis with three possible outcomes: the estimated slope is positive (i.e., the CL of the first derivative contains only positive values), or negative (the CL contain only negative values), or possible zero (the CL contains zero). The SiZer map considers all reasonable bandwidth values and exploits the notion that various values provide different information about the data. Thus SiZer

map displays this information in one image. While SiZer map is constructed by simultaneously fitting a family of smooths, a reasonable statistical inference accounting for multiple comparison testing was addressed by Hannig & Marron (2006) with the advanced distribution theory for SiZer. In this paper, the appropriate quantile  $q(\lambda)$  was used for the multiple testing adjustment (Hannig & Marron, 2006). The LL-SiZer map is generated using r-package ‘‘SiZer’’ (Sonderegger, 2012) which considers the appropriate ‘row-wise adjustment’ (Hannig & Marron, 2006) to compute the critical value of  $q(\lambda)$  in the construction of the CL. For details about other implementation, see Hannig & Marron (2006).

#### 4.2.2 SS-SiZer

The SS-SiZer (Marron & Zhang, 2005) is based on the smoothing spline estimation. SS-SiZer incorporates the smoothing spline model and estimates the regression function by minimizing over functions  $g$ .

$$\{y_i - g(x_i)\}^2 + \lambda \int \{g''(x)\}^2 dx$$

where  $\lambda$  is the smoothing spline parameter that determines the smoothness of the regression estimate  $\hat{g}_\lambda(x)$  and  $\int \{g''(x)\}^2 dx$  represents the roughness of the underlying function  $g(x)$ . According to Green & Silverman (1994), that the solution of the minimizing  $\hat{g}_\lambda(x)$  is a natural cubic spline with knots at the data locations  $x_1 \dots x_n$ .

Smoothing spline SiZer (Marron & Zhang, 2005) uses the ‘independent block’ idea to construct point wise CL to produce the map. In our research, we apply the simultaneous CL to the SS-SiZer model to address the multiplicity comparison issue (same approach as PS-SiZer in Section 4.3.3). The interpretation of a SS-SiZer map remains the same as in

Marron & Zhang (2005). For other implementation details, such as the expression of first derivative estimate, and its standard error, see Marron & Zhang (2005).

#### 4.3 PS-SiZer method

In this section, we propose an extension of the SiZer map in order to handle data that arise in a longitudinal setting. In the proposed method, we consider subject-specific correlation that is inherent from repeated measurement data. In the underlying PS-SiZer model, we consider an approach similar to the standard SiZer model in which a family of smooth functions is used at various levels of smoothing parameters. In our research, we applied penalized spline regression (PSR) as the underlying model as the computationally efficient smoothing model for the following reasons. First of all, SiZer is a powerful exploratory data analysis tool to explore the large HIV data to find the underlying features of weight trajectory. Second, for the large HIV data with more than 2 million observations, PSR model serves as a computationally efficient smoothing method. Third, the PSR model can deal with the subject specific correlation arises from the longitudinal setting. In the development of the PS-SiZer map, we applied the simultaneous CL to resolve the multiplicity comparison issue. The proposed PS- SiZer extends the SiZer as follows:

- 1) Addition of a random intercept component to consider the subject-specific correlation
- 2) Application of P-spline (Eilers & Marx, 1996) as the underlying smoothing function
- 3) Construction of simultaneous 95% CL addressing the multiple comparison issue

#### PS-SiZer model specification

In this paper, we define our underlying model with three components



- 1) A random component
- 2) A smooth function
- 3) An error term

Let  $y_{ij}$  be the outcome measurement on subject  $i$ ,  $i = 1, 2, \dots, n$  at time  $x_{ij}$ ,  $j = 1, 2, \dots, n_i$ . A model for these data is represented as

$$y_{ij} = g(x_{ij}) + b_i + \varepsilon_{ij}; \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2); \quad b_i \sim N(0, \sigma_b^2) \quad (4.1)$$

where  $g(\cdot)$  is a penalized spline regression (PSR) model. The random component  $b_i$  is a random intercept with variance  $\sigma_b^2$ , such that  $b_i \sim N(0, \sigma_b^2)$  and  $\varepsilon_{ij}$  are random errors with variance  $\sigma_\varepsilon^2$  and  $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$  and the  $b_i$ 's and  $\varepsilon_{ij}$ 's are mutually independent.

In this paper, we utilize the P-spline method of Eilers & Marx ((1996), to estimate the underlying population regression function  $g(x_{ij})$ . The P-spline model specification considers B-spline as the basis function. Furthermore, P-spline uses evenly spaced knots with the difference penalty applied directly to the parameters to control the ‘wiggleness’ of the function. Let  $B_m(x_{ij}; p)$  denote B-spline basis of degree  $p$  with  $k'$  equal intervals of  $k' + 1$  knots. Hence, the number of B-spline in the regression is  $M = k' + 1 + p$ . The B-spline smooth function is as follows:

$$B(x) = \sum_{m=1}^M a_m B_m(x; p)$$

where  $\{a_m\}, m = 1, \dots, M$  is a vector of coefficients, and  $B_m(x; p)$  is the B-spline basis function of degree  $p$ .

For penalty, P-Spline (Eilers & Marx, 1996) uses base penalty on higher order finite differences,  $\Delta_d^T \Delta_d$ . Therefore, the difference penalty matrix with order  $d$  can be written as,  $a^T \Delta_d^T \Delta_d a$ .

Here,  $\Delta_d$  is a matrix such that  $\Delta_d$  constructs the vector of  $d^{th}$  difference of  $a$  i.e.,  $\Delta a_m = a_m - a_{m-1}$ ;  $\Delta_2 a_m = a_m - 2a_{m-1} + a_{m-2}$  and so on). For example, the difference matrix of the second order  $\Delta_2$  for 5 coefficients “a1... a5” has the form

$$\Delta_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix}$$

The second component of the model (4.1) is the subject-specific random effect  $b_i \sim N(0, \sigma_b^2)$ .

Therefore, the penalized least square objective function minimizes

$$\|y - Ba - Zb\|^2 + \lambda a^T \Delta_d^T \Delta_d a + (\sigma_\varepsilon^2 / \sigma_b^2) b^T b \quad (4.2)$$

$$\text{where } Z = \begin{pmatrix} 1_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1_n \end{pmatrix} \text{ and } 1_i = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n_i \times 1}$$

The above minimization problem can be handled using the mixed model framework. Equation (4.2) can be turned into a regular mixed model by making use of the mixed effect model framework discussed in details in Brumback et al. (1999), Ruppert et al. (2003), Durban et al. (2005), and Wood (2006b) among others. In this case the minimization of penalized least squares criterion with the associated penalty, is equivalent to maximizing the log-likelihood which arises from  $a$  and  $b_i$ . Here  $a$  and  $b_i$  are treated as a pair of independent random vectors but  $a$  has an improper distribution. The improper distribution for  $a$  does not fit easily into standard linear mixed modeling approaches

(Pinheiro & Bates, 2000). Some re-parameterization is needed. So that the new parameters are divided into a set with a proper distribution, to be treated as random effects, and a set with improper uniform distribution, to be treated as fixed effects (Wood S. N., 2006b). We discuss a mixed model approach for (4.2) in the next section.

### Mixed model representation

Let us first consider the difference matrix  $\Delta_d$  that has the dimensions  $(k' + 1 + p) \times (k' + 1 + p - d)$ . The penalty matrix  $\Delta_d^T \Delta_d$  is singular, and has the rank  $k' + 1 + p - d$ . An eigen value decomposition of  $\Delta_d^T \Delta_d$  leads to  $\Delta_d^T \Delta_d = U \text{diag}(\Lambda) U^T$  with  $U$  as the eigenvectors and  $\Lambda$  is the diagonal matrix of eigenvalues in non-increasing order. Therefore,  $k' + 1 + p - d$  eigenvalues are strictly positive and the remaining  $d$  are zeros.

Hence,  $U$  and  $\Lambda$  can be represented as,  $U = [U_+, U_0]$  and  $\Lambda = (\Lambda_+^T, 0_d^T)^T$ . The dimension of  $U_+$  is now  $(k' + 1 + p) \times (k' + 1 + p - d)$  with corresponding non-zero elements of vector  $\Lambda$ .

Hence, we can rewrite  $Ba$  as

$$\begin{aligned} Ba &= BUU^T a = B \left[ U_0 U_0^T a + U_+ \text{diag} \left( \Lambda_+^{-\frac{1}{2}} \right) \text{diag} \left( \Lambda_+^{\frac{1}{2}} \right) U_+^T a \right] \\ &=: B \left[ U_0 \beta + U_+ \text{diag} \left( \Lambda_+^{-\frac{1}{2}} \right) u \right] =: X\beta + Z_B u \end{aligned}$$

$$\begin{aligned} \text{and, } a^T \Delta_d^T \Delta_d a &= a^T U \text{diag}(\Lambda) U^T a = a^T U_0 \text{diag}(0_d) U_0^T a + \\ & a^T U_+ \text{diag}(\Lambda_+) U_+^T a = u^T u \end{aligned}$$

The mixed model representation of the smooth function is:

$$X\beta + Z_B u \quad \text{where } u \sim N(0, \sigma_u^2 I_{k+1+p-d})$$

And our final model including random intercept takes the form,

$$Y = X\beta + Z_B u + b_i + \varepsilon$$

$$\text{where } u \sim N(0, \sigma_u^2 I_{k+1+p-d}); b_i \sim N(0, \sigma_b^2); \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n) \quad (4.3)$$

Therefore, the model predictor is made up of three components. The first component  $X\beta$  represents the fixed overall effect. The second component,  $Z_B u$  is the smoothing function and the third or random component,  $b_i$ , measures the random departure of the subjects from the overall effect. The estimates of parameters and random coefficients are obtained as the BLUP from mixed model using the REML criteria for variance components. Equation (4.3) can be solved using any standard mixed model software. In our research, we utilized R-package `mgcv::gam` (Wood S. , 2010) to take the computational advantages in fitting equation (4.3). We obtained the estimate of  $g(x)$ , the mean population curve at  $x$  and the quantities of interest to generate the PS-SiZer map. The most crucial component in the PS-SiZer map is to estimate the first derivatives of the fitted functions  $\hat{g}_\lambda(x)$  (i.e.  $\hat{g}_\lambda'(x)$  and the variance of  $\hat{g}_\lambda'(x)$  and associated confidence bands).

#### 4.3.1 Inference

In the previous sections the point estimate for the smoothing model parameters were discussed, yet we are interested in finding the confidence intervals for the smoothing parameters and quantities derived from them, such as estimate of smooth function,  $\hat{g}_\lambda(x)$  and the first derivatives of the smooth function,  $\hat{g}_\lambda'(x)$ . Simon N. Wood (2006b) detailed the formulation of the covariance matrix for the smoothing parameters. In this paper, we discuss how we follow the estimate of the covariance matrix for the

smoothing parameters specified by Simon N. Wood (2006b). Let  $\Phi = \begin{bmatrix} \beta \\ u \end{bmatrix}$  contain all the fixed effects and the random effects from the smooth term only and let  $C = [X \ Z_B]$  be the corresponding model matrix. Let  $Z$  be the random effect model matrix excluding the columns related to smooth terms and  $\sigma_b^2$  be the corresponding random effects covariance. So, the covariance matrix is  $V = \sigma_b^2 Z Z^T + \sigma_\epsilon^2 I$ . Therefore, the estimated covariance matrix ( $\Sigma$ ) for the parameters:

$$\Sigma = cov(\Phi) = (C^T V^{-1} C + \check{D})^{-1}$$

where  $\check{D}$  is the positive semi-definite matrix of the coefficients for the smooth terms.

The standard deviation of the smooth function at point “x”,  $\widehat{SD}(\hat{g}_\lambda(x)) = \sqrt{C_x(\Sigma) C_x^T}$  with  $C_x = [X_x \ Z_{Bx}]$ .

#### 4.3.2 Estimate and variability bands of the derivatives

The derivatives of the smooth function are obtained by defining  $C'_x = [X'_x \ Z'_{Bx}]$ . Here,  $X'_x = \frac{d}{dx}(X)$  and  $Z'_{Bx} = \frac{d}{dx}(Z_x)$ . The estimated first derivative is:

$$\widehat{g}'_\lambda(x) = C'_x \Phi$$

The estimated standard deviation is,  $\widehat{SD}(\widehat{g}'_\lambda(x)) \cong \sqrt{C'_x(\Sigma) C_x^T}$

#### 4.3.3 Confidence band

The construction of the PS-SiZer map involves a family of smooths based on the confidence bands of the derivatives  $\widehat{g}'_\lambda(x)$ . We generated 100 levels of smoothing parameters to construct a PS-SiZer map. The range of the smoothing parameters  $\lambda_{min}$ ,  $\lambda_{max}$  is defined such that  $(\lambda_{min}, \lambda_{max}) = [(\log_{10}(\lambda_{REML}) - 2, \log_{10}(\lambda_{REML}) + 2)]$ . The number 2 is arbitrary, but we chose it to get a reasonable wide range of smoothing

parameter, and we obtained the  $\lambda_{REML}$  from the REML estimate of smoothing parameter using the same P-spline smoothing function.

PS-SiZer can be viewed as a collective summary of a large number of hypothesis testing, and a reasonable statistical inference is necessary for the multiple testing issue. Likewise, Ruppert, Wand, & Carroll (2003) stated that penalized spline has a fairly straightforward simulation based simultaneous confidence bands which can be used in situations when multiplicity testing is carried out. Suppose we want a simultaneous confidence band for  $g(\cdot)$  on a grid of x-values,  $x_{grid} = (x_1, \dots, x_r)$  and define

$$g(x_{grid}) = \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_r) \end{bmatrix}$$

A 100  $(1 - \alpha)\%$  simultaneous confidence band for  $g_\lambda(x_{grid})$  is

$$\hat{g}_\lambda(x_{grid}) \pm q_{1-\alpha}(\lambda) \begin{bmatrix} \widehat{SD}\{\hat{g}_\lambda(x_1) - g(x_1)\} \\ \vdots \\ \widehat{SD}\{\hat{g}_\lambda(x_r) - g(x_r)\} \end{bmatrix}$$

where  $\hat{g}_\lambda(x_{grid})$  be the corresponding EBLUP obtained from mixed model framework.

Here,  $q_{1-\alpha}(\lambda)$  is the  $(1 - \alpha)$  quantile of the random variable at a smoothing level  $\lambda$ ,

$$\sup_{x \in \mathcal{X}} \left| \frac{\hat{g}_\lambda(x) - g(x)}{\widehat{SD}\{\hat{g}_\lambda(x) - g(x)\}} \right| \quad (4.4)$$

which is the supremum or least upper bound on the set  $\{g(x_{grid}): x \in \mathcal{X}\}$ .

The quantile  $q_{1-\alpha}(\lambda)$  was approximated using N=10,000 simulations. The N simulated values are sorted from smallest to largest, and the one with rank  $(1 - \alpha)N$  is used as  $q_{1-\alpha}(\lambda)$ .

For a PS-SiZer map, we obtained the 95% quantile of equation (4.4) based on a simulation of size N at each level of smoothing parameter  $\lambda$ .

Therefore, the confidence limits (CL) are obtained as below:

$$\hat{g}'_{\lambda}(x) \pm q_{1-\alpha}(\lambda) * \widehat{sd}(\hat{g}'_{\lambda}(x)) \} \quad (4.5)$$

#### 4.3.4 Construction of color coded PS-SiZer map

The PS-SiZer map provides the characteristics (i.e. curve increasing, decreasing, or neither) of the estimated curve in the form of a color-coded map. The vertical axis of the PS-SiZer map corresponds to the level of smoothing  $\lambda$ , and the horizontal axis of the PS-SiZer map represents time. At each time-point, PS-SiZer uses a color that indicates inference about the estimated function by means of their corresponding first derivatives (slope):

- 1) The blue color indicates that the smooth function is significantly increasing corresponding to the 95% CL of the slope fully above zero.
- 2) The red color appears in the PS-SiZer map when the smooth function is significantly decreasing corresponding to the 95% CL of the slope fully below zero.
- 3) The purple color is used when there is no significant change corresponding to the 95% CL of the slope containing zero.

#### 4.4 Simulation study

In practice, the fundamental function of a SiZer map is to detect the underlying features of the underlying curve from which the data are generated. For this reason, it is natural to compare the different versions of the SiZer maps with respect to the correct number of underlying features detected.

We have conducted Monte Carlo simulation studies to evaluate the performance of PS-SiZer map under various scenarios. The key objective of this simulation study was to compare the PS-SiZer with the LL-SiZer and SS-SiZer. But in a sense, comparing the curves between methods is not straightforward, because each curve is parameterized differently. Kernel regression (LL-SiZer) would give us an error curve over the bandwidth; smoothing splines (SS-SiZer) or penalized smoothing spline (PS-SiZer) would give us an error curve over the smoothing parameter. So to compare the three SiZer maps, the notion of degrees of freedom gives us a way of precisely making this comparison. The degrees of freedom of a fitting procedure describes the effective number of parameters used by this procedure, and hence provides a quantitative measure of estimator complexity. For this reason all three SiZer maps (PS-SiZer with LL-SiZer and SS-SiZer) were generated with the same range of EDF.

Our simulation study was designed to mimic the HIV data. Using the simulated data, the performance of the SiZer maps were evaluated using the following approaches:

- 1) Each SiZer map was compared at a similar level of “Effective Degrees of Freedom” (EDF).
- 2) The performance of the three SiZer maps was compared according to which flags correct number of features (increasing, decreasing or stable) of a curve.
- 3) The performance of the three SiZer maps was compared according to which is most sensitive to detect the time point where a curve reaches the maximum.

In this research, performance of the PS-SiZer maps is presented in two different simulation studies: ‘Simulation Study One’ and ‘Simulation Study Two’.



#### 4.4.1 Simulation study-one

The data presented in Figure 4-2 (a) were simulated with the observation times  $x_i$  chosen to be equally spaced in  $[0, 1]$  with

$$g(x_{ij}) = 65 + 25e^{-2.0*x_{ij}} * \sin(5\pi(x_{ij} + 5)) + b_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2); b_i \sim N(0, \sigma_b^2)$$

Here,  $x_{ij}$  denote the time of measurement,  $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$  is a random noise,  $b_i$  is a random intercept,  $b_i \sim N(0, \sigma_b^2)$  and the errors are mutually independent.  $\sin(5\pi(x + 5))$  is a periodic function which has five features. Here, we defined features of a curve when a curve changes its status (increasing, decreasing or stable) from one to another. The quantity,  $25e^{-2.0x}$  is a function to control the spread of the periodic function. In addition, subject-specific data were generated by adding the coefficients in model with a random quantity  $b_i$

We investigated three SiZer maps through the various levels of combination of error variance and subject specific variance, such as the ratio  $\sigma_\varepsilon^2 : \sigma_b^2 = (2:5)$ ,  $(5:2)$ , and  $(5:15)$ , respectively.

For each scenario of different variance combination, 50 trials were generated. Each of the trials consisted of  $N=100$  subjects and the number of observations per subject was  $n_i = 10$  for  $i = 1, \dots, N$ . For each simulated trial three different SiZer maps were generated at 100 levels of EDF. In Figure 4-2 (b, c, d), we present example SiZer maps from a randomly chosen trial from the 50 trials generated in the simulation scenario with  $\sigma_\varepsilon^2 : \sigma_b^2 = (5:15)$ . We present the SiZer maps with the time depicted on the x-axis and EDF on the y-axis for a better demonstration of the comparisons of the three maps.

### Comparison of feature detection:

We present the average proportions of the features detected in Table 4-1 based on the 50 simulated data sets at various levels of  $\sigma_\varepsilon^2$  and  $\sigma_u^2$ . The true curve has five features. The first three features (increasing to decreasing to increasing) were prominent features and were detected by the three maps most of the time. We were mainly interested if the PS-SiZer can detect the fourth and the fifth feature of the true curve and compare its performance to LL-SiZer and SS-SiZer. When subject-specific variation is small ( $\sigma_b^2 = 2$ ), PS-SiZer detected five features 51% of the time, whereas SS-SiZer and LL-SiZer were able to detect it 2% and 14% of the time, respectively. Four features were detected by the PS-SiZer, SS-SiZer and LL-SiZer 88%, 47% and 34% of the time, respectively. Thus, PS-SiZer detected four features in most cases while the SS-SiZer and LL-SiZer maps results show that fourth feature has been missed for over half of the considered EDFs. When the subject specific variation is  $\sigma_b^2 = 15$ , PS-SiZer was still able to detect the four features almost 68% of the time compared to 40% by SS-SiZer and only 24% by LL-SiZer. Interestingly, the fifth feature was not detected by LL-SiZer at all in this scenario compared to 8% by PS-SiZer and 5% by SS-SiZer. Similar results were seen in studies with  $\sigma_\varepsilon^2 = 2$  and  $\sigma_b^2 = 5$  (Table 4-1).

Table 4-1 Finding Features: Simulation study-1 with varying variability

Variability ( $\sigma_\varepsilon^2 : \sigma_b^2$ )	Features Detected	Proportion by SiZer Maps		
		LL-SiZer	SS-SiZer	PS-SiZer
5.0 : 2.0	Five	14%	2%	51%
	Four	34%	47%	88%
2.0 : 5.0	Five	4%	10%	30%
	Four	32%	62%	85%
5.0 : 15.0	Five	0%	5%	8%
	Four	24%	40%	68%

Note: proportions are estimated based on the 50 simulated data sets.

We examined the three example SiZer maps from the simulation data (Figure 4-2). Three maps, PS-SiZer, SS-SiZer and LL-SiZer, were able to clearly flag the large two features (blue and then red) in the underlying curve. However, LL-SiZer could not flag most of the periodic region (i.e. the 3rd or 4th features as being statistically significant as expected at half of the time). In fact, the small increasing (blue) fifth feature was not detected by LL-SiZer at all at any level of EDF. LL-SiZer could not detect the small jumps at most of the cases in our simulation study. The three SiZer maps are compared at the same EDF level, for example at  $EDF = 5$  in Figure 4-2. LL-SiZer map was able to identify first three features, increasing to decreasing and to increasing (blue/red/blue) at a level where model EDF was 50. But at the same level of EDF, SS-SiZer was only able to detect the first two features, increasing to decreasing, whereas PS-SiZer map was able to detect all five features that existed in the true function. Similarly, we can compare the three maps at a higher or lower level of EDF. At lower level of EDF (i.e. higher level of smoothing),

SS- SiZer performed better than the LL-SiZer. On the other hand, PS-SiZer performed similarly or better than both LL-SiZer and SS-SiZer at any level.

Therefore, it showed that the first two or three prominent features were detected at wide range of EDF levels by all three approaches. Both of the smoothing spline models (SS-SiZer and PS-SiZer) were somewhat sensitive to the small jumps or features compared to the local linear regression (LL-SiZer) approach. The simulation studies demonstrated that, at a wide range of EDF levels, PS-SiZer was sensitive to even small features at a trivial bump.

We conclude here that the maps from LL-SiZer and SS-SiZer performed similarly and one is not better than the other for all the scenarios considered. The model performance, as well as flagging more underlying features, revealed that the PS-SiZer map is an improved addition to the SiZer map family, especially in longitudinal data analysis.

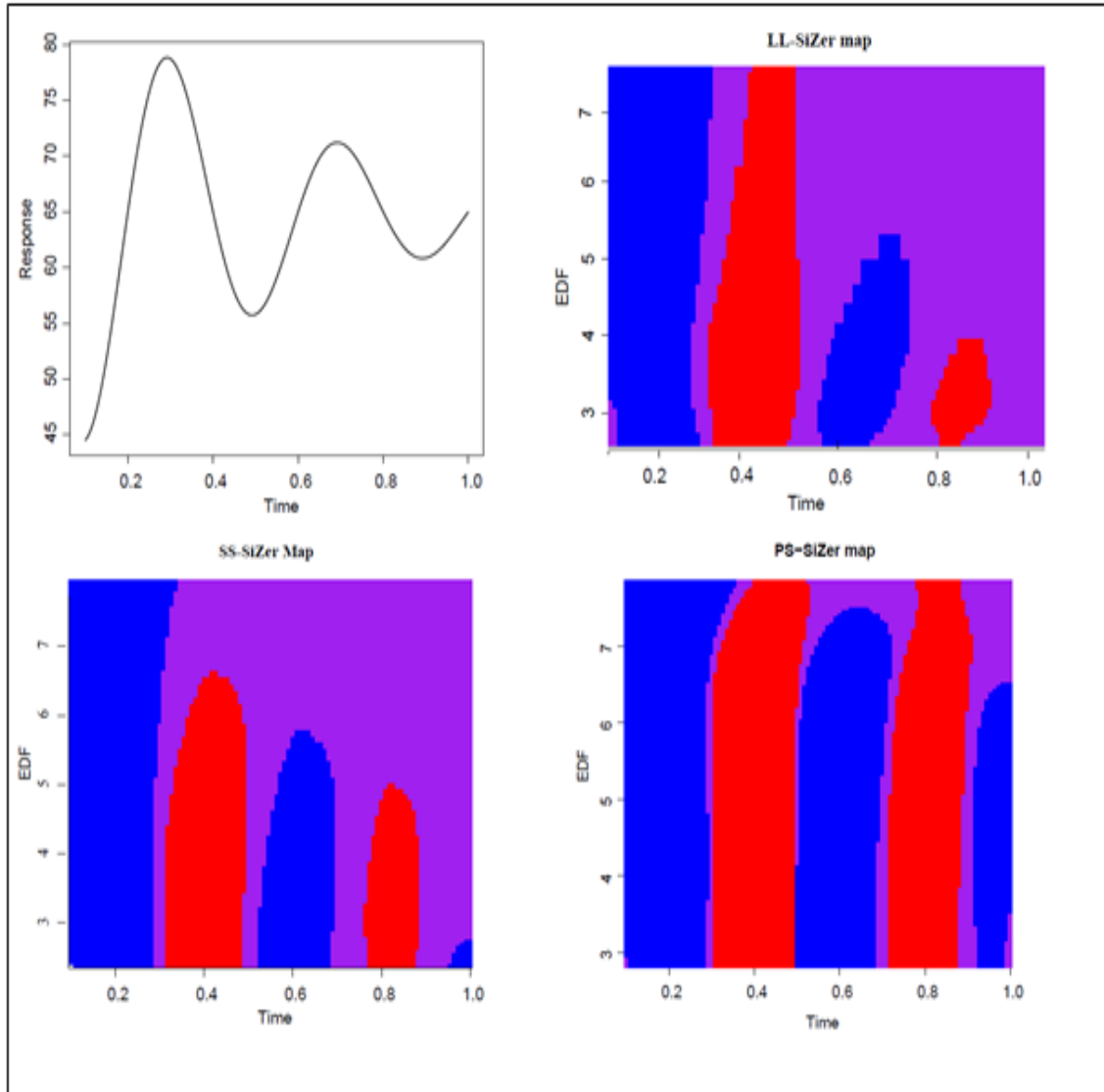


Figure 4-2 Simulation study-one. Upper-left panel: True function where the vertical axis represents responses, and the horizontal axis represents the time; Upper-right panel: LL-SiZer map. Lower-left panel: A SS-SiZer map. Lower-right panel: A PS-SiZer map. For the SiZer map presentation, the vertical axis represents the 100 levels of EDF, and the horizontal axis represents the time.

#### 4.4.2 Simulation study-two

In this simulation study, our aim was to illustrate how sensitive PS-SiZer map is compared to LL-SiZer and SS-SiZer in detecting the point where an increasing function

reaches its maximum. The true curve and the first derivative of the curve are presented in the top left panel of Figure 4.3. The data were generated as  $x_i$  equally spaced in [1:20] with

$$g(x_{ij}) = 85 - \frac{x_{ij}}{4} - e^{(-x_{ij}+4.5)} + b_i + \varepsilon_{ij}$$

where

$x_{ij}$  = Time measurement

$\varepsilon_{ij}$  = Independent random noise such that  $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$

$b_i$  = Subject-specific random intercept with  $b \sim N(0, \sigma_b^2)$

Similar to simulation study-one, we generated 50 data sets, each with  $N = 100$  subjects and 10 equally spaced time points, (i.e.  $n_i = 10, i = 1, 2, \dots, N$ ). The study presented in this paper considered the simulation scenario with the error variance,  $\sigma_\varepsilon^2 = 10$ , and subject-specific variance,  $\sigma_b^2 = 5$ . For each simulated dataset, three SiZer maps were generated at 100 levels of EDF.

This simulation study helped to identify the sensitivity of the SiZer map to detect the point where the increasing curve reaches its maximum. The function used in this example had its maximum at  $x = 4.5 - \ln\left(\frac{1}{4}\right) \sim 5.89$ . The sensitivity of the SiZer maps was calculated at each level of EDF by following:

- 1) For all three SiZer maps, the first time point, where map moves from blue region to purple region, has been calculated at each level of EDF. The process has been repeated for 50 simulation trials. The summary of the time point where the maximum of the increasing curve was detected by the three SiZer maps is presented in a boxplot (Figure 4-3).

The boxplot summary shows that PS-SiZer map is the most sensitive and detects the time point where curve reaches its maximum ( $x \sim 5.89$ ). We presented three example SiZer maps, LL-SiZer, SS-SiZer and PS-SiZer, from a random simulation in Figure 4-4. Three maps were able to detect the pattern of the curve by moving from the blue region to the purple region at all levels of EDF. However, LL-SiZer and SS-SiZer had more blue area on the left than the PS-SiZer map, indicating the maximum of the curve is obtained at a time greater than 6.

Simulation study-two provided a robust indication that the PS-SiZer map not only detects the significant change of the true curve, but also is sensitive enough to detect the true time point where the curve reaches its maximum. Even though all three SiZer maps were able to detect the pattern of the true curve, (that is, the trajectory of the curve from significantly increasing (blue area) to non-significant (purple area) change at almost all levels of the EDF) the LL-SiZer and SS-SiZer were less sensitive in locating the true time point where the curve reaches its maximum compared to PS-SiZer map. If the research interest is to find the time point where the changes occur, then PS-SiZer map will be more sensitive to detect the time of true change competently.

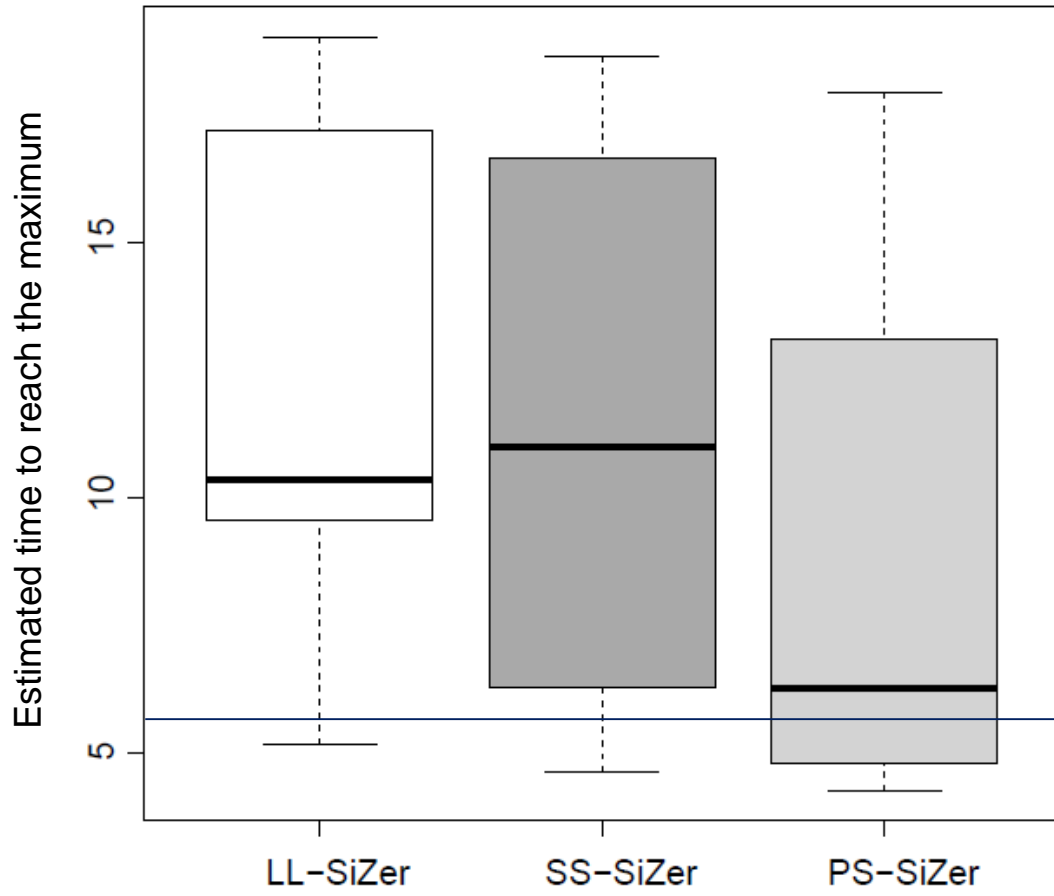


Figure 4-3 Boxplot-Summary of three SiZer maps: Time to detect maximum value.



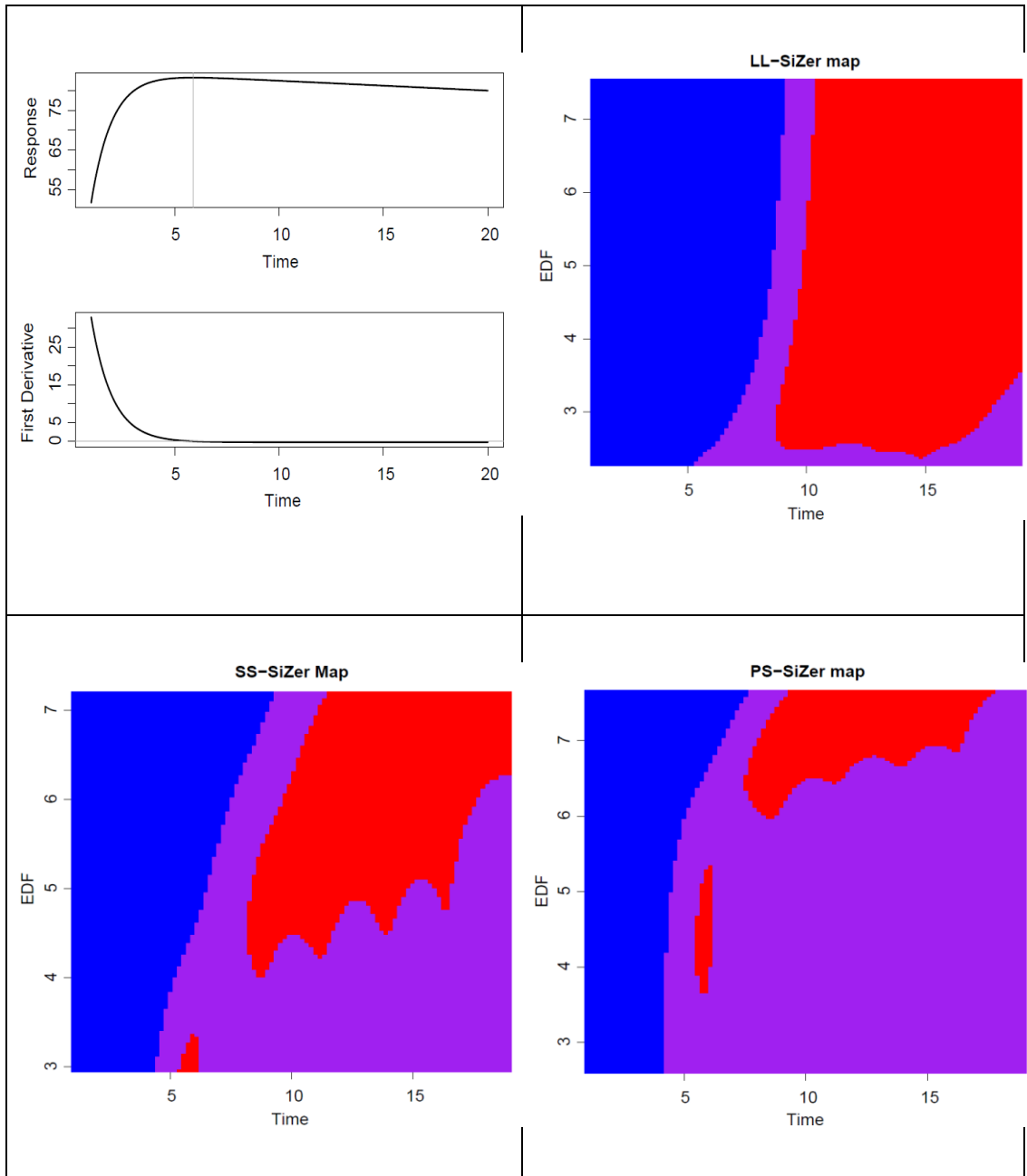


Figure 4-4 Simulation study-two. Upper-left panel: True function and the first derivatives where the vertical axis represents responses, and the horizontal axis represents the time Upper-right panel: A LL-SiZer map. Lower-left panel: A SS-SiZer map. Lower-right panel: A PS-SiZer map. For the SiZer map presentation, the vertical axis represents the 100 level of EDF and the horizontal axis represents the time.

#### 4.5 Application of PS-SiZer map

Statistical analyses were performed using SAS Software 9.3 and R software (2.13.2). SAS was used to create the analysis datasets for each of the five IeDEA regions. The user defined R-functions and the R package SiZer (Sonderregger, 2012) was used to generate LL-SiZer maps. To generate SS-SiZer and PS-SiZer maps, user defined R-functions and the R package mgcv::gam (Wood S. , 2010) was used.

##### IeDEA study

In this study, patients received care in a number of regions taking part in the International Epidemiologic Databases to assess AIDS (IeDEA), a worldwide collaboration of HIV clinical cohorts in five world regions. Namely, the regions are the East-Africa, Southern-Africa, Central-Africa, West-Africa, and Asia-Pacific regions (Egger et al., 2012). Individuals included in this analysis were at least 18 years old at ART-initiation and had at least one weight observation at ART initiation (baseline) and post-ART observation within the first four years of treatment. Body weight was measured in patients participating in the stavudine (d4T) containing regimen as well as in the non-d4T regimen after their initiation.

##### 4.5.1 Results from PS-SiZer

The PS-SiZer maps provided a family of smoothed curves considering the subject-specific correlation. PS-SiZer maps were generated for IeDEA regions whose data were analyzed in the present study. PS-SiZer maps generated based on the data collected in the East Africa IeDEA region are illustrated in Figure 4-5, for d4T-containing (a1) and non-d4T-containing first-line ART regimens (a2).

The vertical axis represents the level of smoothing, and the horizontal axis represents the time in weeks since the start of ART. For example, for the d4T regimen, the rate of body weight change shows a significant increase around 50 weeks at most of the scale (-2, 1) because of the blue area on the left of the map. The “purple” on the right indicates there is no more significant increase or decrease after around weeks 50. There is a fairly wide range of scale (-1, 1) where the red area indicates a significant weight decrease between 60 weeks to ~90 weeks. The blue area in the map from weeks 150 and beyond indicate a significant increase at a fairly wide range of scale (-1.5 to 0.5). Similarly, at the very high scale (>1.0), the entire region over time is blue, mostly because of over-smoothing.

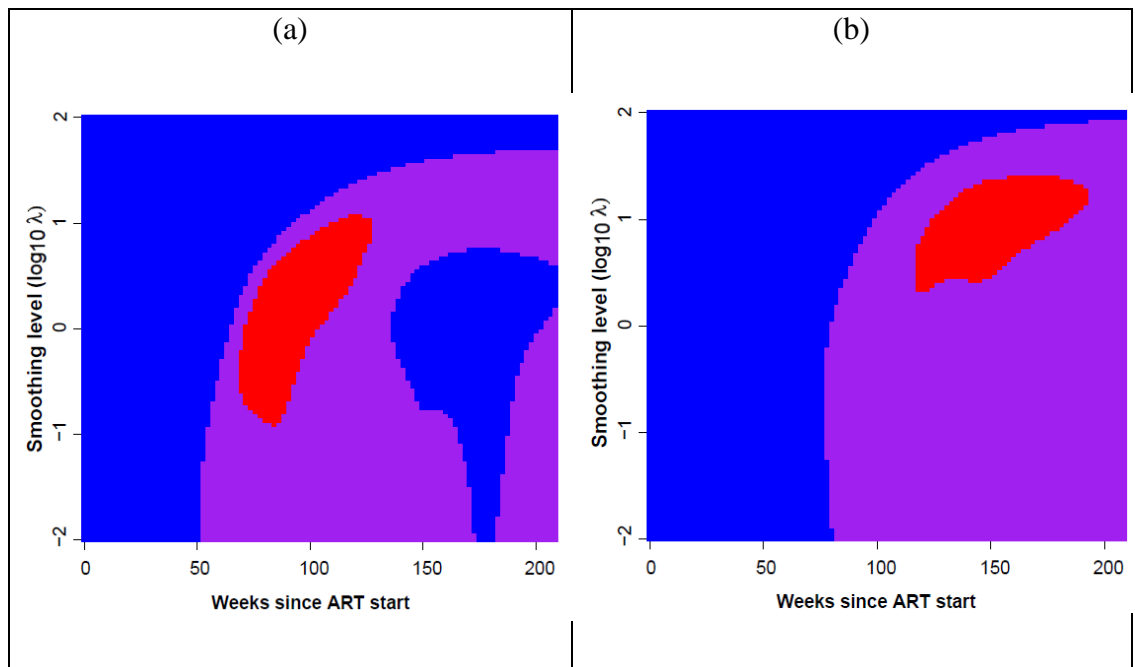


Figure 4-5 PS-SiZer map from East Africa region. (a) Left panel represents SiZer map of rate of body-weight changes over time (weeks since ART start) for patients initiated with d4T-treatment regimen. (b) Right panel represents SiZer map of rate of body-weight changes for patients initiated with non-d4T-treatment regimen. The vertical axis represents the level of smoothing,  $\lambda$  and expressed as  $\log_{10}(\lambda)$ , and the horizontal axis represents the weeks since the start of ART treatment.

On the other hand, the PS-SiZer map of rate of bodyweight change for HIV patients initiated with the non-d4T regimen from the same region illustrates relatively consistent significant increases until about week 90. This indicates that for non-d4T treated patients, the rate of bodyweight significantly increases almost up to 2 years, but there is a trend to have significant decreasing weight after 2 years on treatment. PS-SiZer analyses were generated for each of the remaining four IeDEA regions and are added in Chapter 5.

#### 4.6 Discussion and conclusion

From a technical perspective, the standard LL-SiZer and SS-SiZer methods do not consider the correlation induced by repeated measurements obtained in the same patient, which invariably arise in longitudinal settings. Our proposed PS-SiZer model was developed by keeping that limitation in mind. Our key efforts were centered on developing SiZer maps for correlated data with appropriate and computationally efficient smoothing methods and resolving this correlation problem.

In simulation studies, the PS-SiZer map outperformed both LL-SiZer and SS-SiZer in the analysis of data arising from a longitudinal setting. The fundamental motivation of a SiZer map is to detect the underlying features in the data. Therefore, the key goal of this research was to show which SiZer map could detect correct number of features in longitudinal settings. From the simulation results, it was evident that both standard LL-SiZer and SS-SiZer methods clearly flag the large features in the data. However, LL-SiZer could not flag most of the periodic region as being statistically significant as expected most of the time. Similarly, SS-SiZer, which uses a smoothing spline estimate without consideration of correlation, was somewhat less sensitive to the small jumps or features. The prominent jumps were detected by SS-SiZer at small to large EDF levels, but it could

not detect some features at larger EDF levels compared to LL-SiZer. Marron & Zhang (2005) also attempted to compare these two maps with various simulations studies. The authors concluded that the original local linear version (here, LL-SiZer) of SiZer and smoothing spline SiZer (here, SS-SiZer) often performed similarly, but they could not conclude that one is always better than the other. Similar findings were observed in our simulation studies.

The PS-SiZer maps from the proposed approach flagged more underlying features in the simulation data. At a wide range of EDF (low to high), PS-SiZer consistently detected underlying features better than other two SiZer maps. The simulation studies demonstrated that at a moderate range of EDF levels, PS-SiZer was sensitive to small features, even for a trivial bump. In summary, PS-SiZer deserves the merit to be a new addition to the existing family of SiZer maps to enhance the analysis of data from longitudinal settings.

The main idea of SiZer maps is to detect significant changes by mapping areas where the 95% confidence intervals of the first derivative is above zero (significantly increasing), below zero (significantly decreasing), or contains zero (no significant change). For this reason, the precise estimation of 95% confidence intervals of the first derivative is important and will enhance the detection capability of features in SiZer analysis. The use of the PSR model with random intercepts in PS-SiZer map results in narrower confidence intervals, which, in turn, result in a more accurate detection of features compared to standard SiZer maps. The proposed PS-SiZer maps considered the estimation of the variability that was inherent from the study design and, thus, detected features precisely.

## REFERENCES

- Brumback, B. A., & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93(443), 961-976.
- Brumback, B., Ruppert, D., & Wand, M. (1999). Comment on ‘Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, 94:794–797.
- Chaudhuri, P., & Marron, J. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447), 807-823.
- Currie, I. D., & Durban, M. (2001). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 2(4), 333-349.
- De Boor, C. (1977). Package for calculating with B-splines. *SIAM Journal on Numerical Analysis*, 14(3), 441-472.
- De Boor, C. (1978). *A practical guide to splines*. New York: Springer-Verlag, revised edition.
- Durban, M., Harezlak, J., Wand, M., & Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24:1153–1167.
- Egger, M., Ekouevi, D., Williams, C., Lyamuya, R., Mukumbi, H., Braitstein, P., Woolson-Kaloustian, K. (2012). Cohort Profile: The international epidemiological databases to evaluate AIDS (IeDEA) in sub-Saharan Africa. *Int J Epidemiol*, 41:1256–64.
- Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

- Eubank, R. L. (2000). *Spline regression, Smoothing and Regression: Approaches, Computation, and Application*. Wiley Online Library.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.
- Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press. CRC press: Chapman & Hall.
- Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. London: Chapman and Hall.
- Hannig, J., & Marron, J. (2006). Advanced Distribution Theory for SiZer. *Journal of the American Statistical Association*, 484–499.
- Hardle, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University press.
- IeDEA. (n.d.). Retrieved from International Epidemiologic Databases to Evaluate AIDS: <http://www.iedea.org/>
- Loader, C. (1999). *Local regression and likelihood*. New York: springer.
- Marron, J. S. (1996). A Personal View of Smoothing and Statistics in Statistical Theory and Computational Aspects of Smoothing eds. *Haerdle and Schimek, 1-9. Physica-Verlag*, 103-112.
- Marron, J., & Zhang, J. (2005). SiZer for smoothing splines. *Computational Statistics*, 20:481–502.
- Marx, B., & Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 193-209.

- Park, C., & Kang, K. (2008). SiZer analysis for the comparison of regression curves. *Computational Statistics and Data Analysis*, 3954–3970.
- Park, C., Hannig, J., & Kang, K. (2009). Improved SiZer for time series. *Statistica Sinica*, 1511 – 1530.
- Park, C., Marron, J., & Rondonotti, V. (2004). Dependent SiZer: goodness-of-fit tests for time series models. *Journal of Applied Statistics*, 999–1017. .
- Pinheiro, J., & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.
- Rondonotti, V., Marron, J., & Park, C. (2007). SiZer for Time Series: A New Approach to the Analysis of Trends. *Electronic Journal of Statistics*, 268–289.
- Ruppert, D., Wand, M. P., & Carroll, R. (2003). *Semiparametric Regression*. New York:: Cambridge University Press.
- Sonderegger, D. (2012). *SiZer: Significant Zero Crossings* . Retrieved from <http://cran.r-project.org/web/packages/SiZer/SiZer.pdf>
- Staniswalis, J., & Lee, J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93, 1403–1418.
- Wahba, G. (1991). *Spline models for observational data*. Philadelphia: Vol. 59. Siam.
- Wand, M. P., & Jones, M. (1995). *Kernel Smoothing*. London: Chapman and Hall.
- Wood, S. (2010). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*. Retrieved from <http://cran.r-project.org/package=mgcv>.
- Wood, S. N. (2006b). *Generalized Additive Models: An Introduction with R* . Chapman and Hall/CRC.



CHAPTER 5. AN APPLICATION OF PS-SIZER MAP TO INVESTIGATE  
SIGNIFICANT FEATURES OF THE RATE OF CHANGE OF BODY-WEIGHT  
PROFILE FOR HIV INFECTED PATIENTS IN IEDEA STUDY

Abstract

Objectives: Our work involves standardized data collected on HIV-positive patients initiating antiretroviral therapy (ART) in five regions of the International Epidemiologic Databases to Evaluate AIDS (IeDEA) collaboration. The key objective is to understand the pattern of body weight change in HIV patients initiating stavudine (d4T) containing first-line regimens versus non-d4T-containing regimens. This methodology can be adapted to address questions for the evolution of longitudinally collected biomarkers in similar contexts.

Methods: Penalized Spline Significant Zero Crossings of Derivatives (PS-SiZer) is a powerful graphical tool for exploring structures in curves by mapping areas where rate of change is significantly increasing, decreasing, or does not change. In our research, we applied PS-SiZer, an extension of SiZer (Chaudhuri & Marron, 1999) to take into account the within-subject correlation. In the present context, PS-SiZer maps provide information about the significant rate of weight change that occurs in two ART regimens at various level of smoothing. Final conclusions are assessed based on the optimal level of the smoothing parameter, chosen automatically via Restricted Maximum Likelihood (REML) using mixed model representation of the penalized spline regression model. By doing so, we compared the durability of weight gain in patients who received ART regimens containing and not containing d4T. Patients with at least one baseline and follow-up body weight measurement within four years after initiation of ART were included in the analysis.

Results: Statistical analyses included 185,010 patients from five IeDEA regions, consisting of Southern Africa (65.6% of the cohort), East Africa (21.9%), West Africa (8.3%), Central Africa (3.2%), and Asia-Pacific (0.9%). We compared patients initiating ART with a d4T (53.1%) versus a non-d4T (46.9%) containing regimen within each region. The largest difference in the durability of weight gain was observed in Southern Africa where the durability of weight gain in patients treated with d4T-containing regimens lasted 59.9 weeks compared to 133.8 weeks for patients starting ART with non-d4T-containing regimens. Results were similar for the other regions, albeit attenuated. The difference between the durability of weight gain for d4T vs. non-d4T containing regimens was around 49 weeks in West Africa, 32 weeks in East Africa, and 16 weeks in Asia-specific. The durability of weight gain in Central Africa was comparatively similar (difference was -1 week). Overall, d4T-containing regimens were associated with a shorter durability of weight gain, lasting between 39-62 weeks versus 55-134 weeks in patients receiving non-d4T-containing regimens.

Discussion: Results from PS-SiZer maps and the smoothing model at the optimum level showed that patients starting ART with d4T-containing regimens experienced weight gains for shorter periods compared to patients receiving non-d4T-containing regimens.

## 5.1 Introduction

Rates of those suffering from HIV/AIDS are especially high in low and middle-income countries (LMIC). For example, despite constituting only 11% of the total Earth's population, the sub-Saharan Africa region is among the world's epicenters of HIV/AIDS. The numbers are daunting. In 2012, it was estimated that over 35 million individuals were living with HIV with the majority of those in sub-Saharan Africa (>25

million) and approximately four million in Southeast Asia (UNAIDS). Nearly 10 million individuals were receiving ART in LMIC) at the end of 2012 (Boulle, et al., 2014). Because of limited resources in LMIC, many areas in these countries do not have lab facilities to support monitoring of HIV treatment. Therefore, researchers must rely on clinical parameters such as weight.

In LMIC, where therapeutic options are limited, it is important to assess the effectiveness of different combinations of ART regimens, particularly those constituting the first-line treatment options provided to patients. Until 2010, a common antiretroviral medication was provided as part of first-line ART was stavudine (d4T), a nuclease reverse transcriptase inhibitor which was in the World Health Organization's (WHO) list of essential medicines (WHO, 2013). Stavudine is no longer part of the WHO 2013 guidelines for first-line regimens due to a number of adverse side effects; however, it is still widely used in the LMIC setting. In particular, regimens containing stavudine (d4T) were widely prescribed medication as part of a first-line ART regimen in a number of LMIC (Rosen, Long, Fox, & Sanne, 2008). Stavudine has played a critical role in the scaling up of combination ART therapy in LMIC. The World Health Organization notes that in 2009, approximately 56% of HIV regimens in LMIC contained d4T.

Despite its wide use, stavudine has been associated with a number of toxicities including serious neurological (Subbaraman, Chaguturu, Mayer, Flanigan, & Kumarasamy, 2007) and metabolic toxicities such as lipodystrophy and lipoatrophy (Joly et al., 2002; Gallant, Staszewski, & Pozniak, 2004); (Gallant, Staszewski, & Pozniak, 2004) both reflecting distribution of fat in the body. Van Griensven (2007) also demonstrated the development of lipoatrophy and subsequent weight loss after initiation

of an ART regimen containing d4T versus zidovudine (AZT)-containing regimen. Another study was conducted by Van Griensven et al. (2010) conducted in which 609 adults were given stavudine for a period of one year in order to assess the weight manifestation after one year of treatment. In about 62% of the patients, weight loss was observed after the first year while no weight gain occurred in any of the participants. It was also concluded that weight loss that was constant and progressive was suggestive of the development of lipotrophy. Stavudine has also been associated with long-term weight loss compared to regimens containing Tenofovir in a randomized study (Gallant, Staszewski, & Pozniak, 2004). They compared Tenofovir DF and d4T regimen in a 3-year randomized trial which showed patients in both treatment groups gained weight during the first 24 weeks. Thereafter, patients who received the d4T regimen progressively lost weight and returned to baseline by week 144 compared to the Tenofovir DF group who showed a stable increase in weight. The overall mitochondrial toxicity was also significantly more among patients receiving d4T regimen.

In programs in LMIC where laboratory access is limited, an ideal way to monitor patients' responses to ART is by measuring clinical parameters such as weight. For example, in developing countries where mortality within the first year of ART is high, early detection of patients with suboptimal ART response (such as weight loss) is crucial. Madec et al. (2009) showed that short-term weight gain is an indicator of treatment success, whereas long-term weight loss is associated with an increased risk of death and other adverse clinical outcomes. Therefore, the durability of weight gain over time can be used as an indicator of the efficacy of the ART treatment in HIV-infected patients (Grinspoon & Mulligan, 2003). Results from a number of studies have shown that weight loss greater

than 10% from ART initiation or the previous visit was significantly associated with a four to six-fold increase in mortality compared with stable or increasing weight (Biadgilign, Reda, & Digafe (2012), Madec, et al. (2009)).

Understanding the prognosis and evolution of HIV disease is important for patient management and assessing the efficacy of the treatment programs. However, which biomarkers to measure remains a challenge, particularly with limited resources in LMIC countries where simple-to-obtain measurements are needed. One attractive option is measuring the body weight of patients with HIV as a follow-up measure (Madec, et al., 2009). Body weight loss is a frequent outcome of infected HIV patients; conversely, weight gain has been seen with the initiation of ART (Wools-Kaloustian, et al., 2006). Increase in weight is an important factor associated with patient survival (Biadgilign, Reda, & Digafe, 2012). WHO guidelines (2003) recommend looking at a body-weight gain of at least 10% at six months from start of antiretroviral therapy to evaluate ART programs. Therefore, the durability of weight gain over time can be used as an indicator of both the efficacy of the ART treatment in HIV-infected patients as well as the efficiency of the treatment program.

The primary objective of our research is motivated by these two, yet still not fully answered, questions: (1) “How does one compare durability of body weight for patients in a d4T regimen vs. a non-d4T regimen in LMIC?” and (2) “How does one assess durability of weight at the time point where weight no longer increases in HIV patients in a d4T regimen vs. a non-d4T in LMIC?”.

To provide a global view of the weight changes after ART initiation (i.e. weight increasing or decreasing), a technique named Penalized Spline Significant Zero Crossings of Derivatives (PS-SiZer), an extension of SiZer (Chaudhuri & Marron, 1999), was

applied. The application of this method to longitudinally collected weight measurements in HIV patients to investigate significant features is novel. An algorithm has been used to estimate the time point where weight no longer increases in HIV-infected patients starting a first-line ART regimen. The algorithm was based on the REML estimate of an optimum level of smoothing using a mixed model representation of the penalized spline regression model (Section 5.2.2). The details of the statistical method of PS-SiZer are presented in Chapter 4, PS-SiZer: A Visual Tool to Investigate Significant Features in Longitudinal data.

## 5.2 Methods

### 5.2.1 Population

For this study, records from patients receiving care at sites participating in the International Epidemiologic Databases to Evaluate AIDS (IeDEA), a collaboration of HIV clinical cohorts representing seven regions of the world, were included in this analysis. The IeDEA Collaboration (Egger, et al., 2012) collects demographic, clinical, and laboratory data extracted from information obtained from patients as part of routine clinical care. The data collected within IeDEA include demographic measures (such as age, sex, geographical region); clinical measures (such as weight, height, medication, morbidity and pregnancy), and biological measures (such as CD4 count, viral load etc.).

The present study includes data on adult HIV-infected patients from five of the seven IeDEA regions representing 140 sites. They include Southern-Africa (87 sites), East-Africa (10 sites), West-Africa (15 sites), Central-Africa (10 sites), and Asia-Pacific regions (18 sites). Individuals included in this analysis were at least 18 years old at ART-initiation

and had at least one weight observation at ART initiation (baseline) and post-ART observation within the first four years of treatment.

### 5.2.2 Statistical analysis

As the longitudinal trajectory of body weight can be rather variable, parametric models may have limitations in detecting features of the regression curve (Jaroslaw, Elena, & Nan, 2007). Therefore, the smoothing curve might be useful for extracting meaningful features in the data. In order to explore the features of the weight trajectory, we advocated using smoothing curve techniques. In our research, we apply PS-SiZer (Chapter 4), the newly developed extension of SiZer methodology for a visual presentation of the overall illustration of the rate of weight changes in HIV-infected patients initiating ART. PS-SiZer is an extension designed to handle correlated data arising from longitudinal settings and an enhancement using computationally the efficient smoothing model. While PS-SiZer map is constructed by simultaneously considering a family of smooth, the ultimate goal is to understand the relationship of the ART treatment regimen (particularly whether the regimen contained d4T or not) and the rate of change of bodyweight over time.

Finally, conclusions were based on the optimum level of the smoothing parameter to obtain the estimated first time point where weight no longer increased for patients in the d4T regimen vs. the non-d4T regimen. Here, we used an algorithm to determine the first time point at which weight gain stops increasing. In the following subsections, the PS-SiZer method and the algorithm are presented briefly.

### 5.2.3 PS-SiZer Maps

PS-SiZer is a visualization tool to investigate significant features in longitudinal data accounting for correlation in the analysis. The details of this method of PS-SiZer are

described in Chapter 4. The underlying smoothing model used in PS-SiZer is a penalized spline regression model (PSR) called P-spline proposed by Eilers & Marx (1996). P-spline is a computationally efficient smoothing model for correlated data.

The PS-SiZer map includes a number of levels of the smoothing parameters. In the present application, we used a range between  $\log_{10}(\lambda_{REML}) \pm 2$ , where  $\lambda_{REML}$  is the estimated smoothing parameter obtained via the REML approach using mixed model representation of the PSR model. One hundred smoothing levels (represented as  $\log_{10} \lambda$  in the map) were used in this analysis to produce the PS-SiZer maps. The PS-SiZer maps focused on the rate of weight change (first derivative) of the curve by smoothing the trajectory of the weight measurements over time. By smoothing the curve, critical patterns in its evolution can be discerned. The significant feature is obtained from the confidence limits (CL) of the first derivatives of the fitted curve at each level of the smoothing parameter. At a specific time point, if the lower limit of CL of the first derivative is above zero, then weight is significantly increasing. When the upper limit of the CL is below zero, the weight is significantly decreasing. The weight is not changing significantly when the CL contains zero. Thus, identifying this point of the derivative crossing zero is critical in estimating the durability of weight increase.

The PS-SiZer analysis explores these characteristics in the form of a color-coded map. Each row of the PS-SiZer map corresponds to a different level of smoothing, and each column represents the weeks from ART initiation. At each time-point, PS-SiZer uses a color that indicates the behavior of the first derivative of the underlying curve:



- 1) The blue color means that the rate of change is significantly positive (i.e., the underlying curve of weight measurements is increasing). In statistical terms, this means that the 95% CL of the rate of change lies completely above zero.
- 2) The red color implies that the rate of change is significantly negative (i.e., the underlying curve of weight measurements is decreasing). This means that the 95% CL of the rate of change is entirely below zero.
- 3) The purple color is used when the CL of the rate of change contains zero, i.e., when there is no significant increase or decrease in the underlying weight measurements over time.

PS-SiZer Maps for each IeDEA region were generated for each treatment group, i.e., one map each for the groups of patients initiating ART with a regimen containing or not containing d4T.

By presenting the features of the underlying weight change at various levels of smoothing, the PS-SiZer map provides an overall visual representation of the weight change after the start of ART. However, to reach a conclusion on the durability of weight increases after ART start, we need to decide on a single optimum level of smoothing. In the below section, an algorithm using a single optimum level from REML estimate is briefly described.

#### 5.2.4 Algorithm to detect first time point (week) at which weight gain stops increasing at an optimum smoothing level

Our algorithm does not depend on a specific smoothing technique. However, here we have used P-spline (Eilers & Marx, 1996) PSR model for its computational efficiency and flexibility for correlated data. In addition, we took the advantages of re-expressing the

PSR model as a linear mixed effect model (Brumback, Ruppert, & Wand, 1999) .The REML estimate of the mixed model is used to obtain the optimum smoothing parameter. Here, the P-spline model includes a subject-specific random intercept to account for correlation arose from repeated weight measures over time. It provides precise estimates of variability and improves the inferences. We used the first derivative of the smooth function to estimate the rate of weight change and the corresponding CL to determine the significant changes over time. Finally, we detect the first-time point of weight change with a 95% CL at an optimum level of smoothing. The steps involved to obtain the first-time point at which weight gain stops increasing at an optimum level of the smoothing are presented below.

- 1) First fit the P-spline model using mixed model representation and obtain the fitted mean regression function and their subject specific deviation at the REML estimated optimum level of smoothing parameter.
- 2) Obtain the estimate of the variance parameters,
- 3) Estimate the first derivatives and their corresponding CL.
- 4) Find the significance feature of the curve from the CL of the estimated first derivatives.
- 5) At each time point, determine if the weight change is significantly increasing if the lower limit of the 95% CL is above zero. Or, determine if the weight change is significantly decreasing if the upper limit of the CL is below zero. It is non-significant if the CL contains zero.
- 6) Find the time point (week) at which the estimated curve stops increasing for the first time.

- 7) Estimate the variance of the first time point using delta method.

Statistical analyses were performed using SAS Software 9.3 and R software (2.13.2) (2008). SAS was used to create the analysis data sets for each of the five IeDEA regions. The user defined R-functions along with the R package `mgcv::gam` (Wood, 2010) was used to generate PS-SiZer maps and to conduct the analysis at the optimum level of smoothing.

### 5.3 Results

#### 5.3.1 Baseline characteristics

The present study includes data on 185,010 adult HIV-infected patients from five of the seven IeDEA regions: Southern Africa (65.6% of the cohort), East Africa (21.9%), West Africa (8.3%), Central Africa (3.2%), and Asia Pacific (0.9%). Total observations at the 4-year follow-up used in the analysis were about 2,161,515. Baseline demographic data of IeDEA patients identified by region and by d4t-containing and non-d4T-containing regimen are shown in Table 5-1.

Table 5-1 Summary of baseline characteristics- IeDEA study by d4T and non-d4T containing regimen

	<b>Asia Pacific</b>	<b>Central Africa</b>	<b>East Africa</b>	<b>Southern Africa</b>	<b>West Africa</b>	<b>Total</b>
<b>d4T Treatment Regimen</b>						
<b>N</b>	963	2839	30990	55192	8176	98160
<b>Female (%)</b>	410 (43)	2008 (70)	20017 (78)	36227 (49)	5490 (55)	64152 (65)
<b>Age (years)</b>	35 (29-40)	37 (31-44)	37 (31-43)	35 (30-42)	39 (32-42)	36 (30-42)
<b>Baseline Body weight (kg)</b>	51 (45-58)	56 (49-65)	54 (48-61)	55 (48-62)	55 (48-64)	55 (48-62)
<b>Baseline Cd4 count (cell/<math>\mu</math>L)</b>	66 (25-153)	127 (57-197)	103 (41-175)	121 (58-190)	138 (58-223)	144 (76-211)
<b>Non-d4T Treatment Regimen</b>						
<b>N</b>	751	3045	9571	66295	7188	86850
<b>Female (%)</b>	181 (24)	2118 (51)	5758 (22)	38137 (51)	4488 (45)	50682 (58)
<b>Age (years)</b>	34 (29-42)	37 (31-44)	37 (31-43)	35 (30-42)	41 (37-42)	36 (30-42)
<b>Baseline Body weight (kg)</b>	58 (50-56)	56 (50-65)	55 (49-62)	55 (49-62)	57 (50-65)	55 (49-62)
<b>Baseline Cd4 count (cell/<math>\mu</math>L)</b>	149 (46-221)	150 (77-223)	109 (44-180)	146 (80-213)	146 (69-224)	144 (76-211)

Note: Summaries are median (IQR) or n (%)

In this research paper, we present in detail results from the Southern Africa IeDEA region. Results from the remaining four regions are presented in less detail as they were, in general, consistent with the results obtained in the Southern-Africa region. When this is not the case, we devote more time to presenting those results.

### 5.3.2 Exploratory data analysis

To clarify what happens to body weight after the initiation of ART, we present, in Figure 5-1 (left-panel), a plot of the individual patient bodyweight measurements over time since the initiation of ART start, with a PSR smoothed curve superimposed to represent the average weight change over time. A random sample of data in 300 patients, taken from the Southern Africa IeDEA region, was used to generate this Figure 5-1. This initial crude representation of the data, nevertheless, points to a sustained increase in weight after ART initiation (with the possibility of a period of weight stagnation or decrease at the rightmost extreme period after ART initiation). To gain some familiarity with these data, we produced a smoothed PSR curve for the weight measurements over time in HIV-infected patients starting ART with d4T versus non-d4T-containing regimens. Data from the Southern African IeDEA region are shown in Figure 5-1 (right panel). From the observed data, it is evident that the pattern of body weight changes among patients treated with the two types of regimens is markedly different.

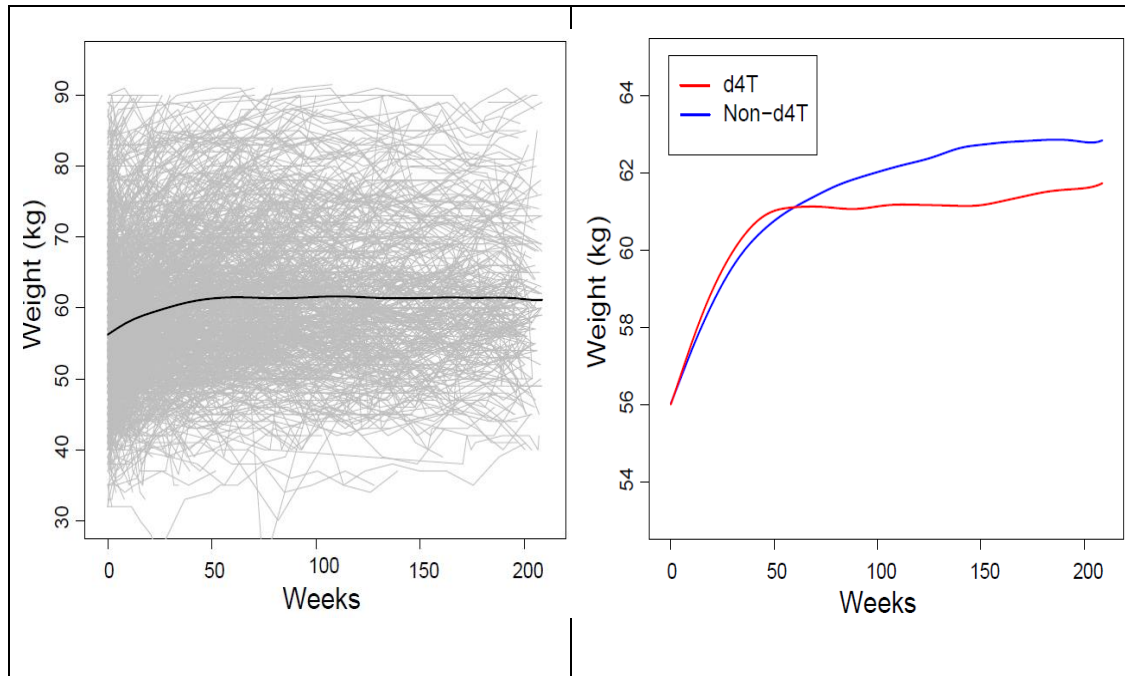


Figure 5-1 Example from the Southern Africa IeDEA region. Left panel: A spaghetti plot of observed weight and P-spline fit over time (weeks). (b) Fitted curves of weight over time by ART treatment regimen using P-spline model.

### 5.3.3 PS-SiZer maps and durability of weight gain at optimum smoothing

PS-SiZer analyses were generated for each of the five IeDEA regions whose data were analyzed in the present study. These are presented in Figures 5-2 to Figure 5-6. Each figure consists of four panels. The smoothed trajectories of weight after ART initiation at the optimum level of smoothing for the two types of regimens are shown in top row: left panel, while the smoothed first derivative of the weight change over time for the two types of regimens is shown in the right panel. The PS-SiZer maps are shown in panels' (bottom row: in the left and right panels) for d4T-containing (left-panel) and non-d4T-containing first-line ART regimens (right-panel).

PS-SiZer maps generated from data in the Southern Africa IeDEA region are illustrated in Figure 5-2. The vertical axis represents the level of smoothing, and the

horizontal axis represents the time, in weeks, since the start of ART as described in the Methods section. For example, for d4T-containing regimens, at a medium level of smoothing (0.5-1.0), body weight increases for about 50 weeks and is reflected by the blue color on the left of the PS-SiZer map. The area to the right of the blue region is colored purple, indicating that no more significant increases in body weight are evident after about 50 weeks from the start of ART. There are some red and blue regions in the map at the lowest smoothing levels (i.e. for values below 0.5), indicating possible weight decreases and increases, respectively. These disappear at higher levels of smoothing. Similarly, at very high smoothing levels, (i.e., for values of the smoothing parameter  $\lambda > 1.0$ ), the entire map is blue, indicating steady weight increases for the entire follow-up period. The PS-SiZer map of bodyweight changes among HIV-infected patients initiating ART with a non-d4T-containing regimen in the Southern African IeDEA region shows that, at lower smoothing levels, there are some blue and purple areas. This suggests an intermittent weight increase. Otherwise, the map consists of mostly blue areas (indicating weight increases) for medium and higher levels of smoothing for up to about 100 weeks after ART initiation. This indicates that patients starting ART with a non-d4T-containing regimen experience sustained bodyweight increases for a period possibly double that of patients treated with d4T-containing regimens.



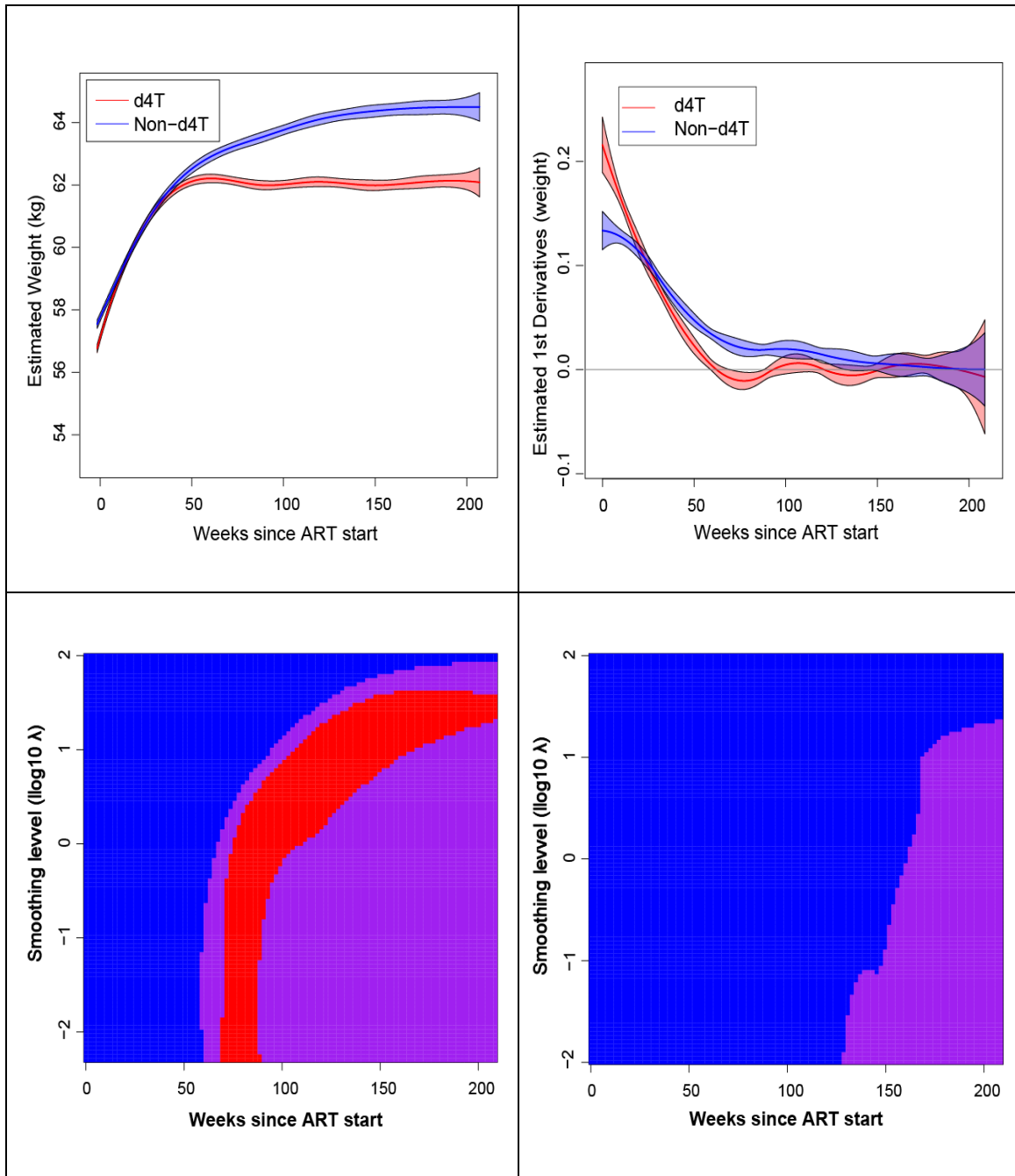


Figure 5-2 Top-row: left panel represents the estimated weight change with CL, and the right panel represents the estimated first derivatives with CL over time by d4T and non-d4T regimens at the REML optimum level. Bottom- row: left panel represents the PS-SiZer map for d4T-regimen, and the right panel represents the PS-SiZer map for non-d4T-regimen. The vertical axis represents the level of smoothing expressed in  $\log_{10} \lambda$  and the horizontal axis represents the time in weeks since the start of ART. Data is from Southern Africa.

To reach a conclusion about the comparison of the durability of weight changes in the d4T-containing versus not-d4T containing regimens, we chose the optimum level of smoothing for the Southern Africa IeDEA region (Figure 5-2: top-row left panel). This analysis showed that weight in patients treated with d4T-containing ART regimens increased rapidly after ART initiation and flattened out afterwards. Consulting the first derivative (Figure 5-2: top-row, right panel), we observed that the 95% CI of the curve includes zero after 59.92 weeks in the group of patients who received a d4T-containing regimen compared to 133.82 weeks for patients treated with non-d4T-containing regimens (panel b2). A numerical summary of these results is also shown in the first row of Table 5-2.

Table 5-2 Estimated weeks at which HIV-Patients experienced non-increasing weight

IeDEA Region	d4T regimen Estimated Weeks (CI)	Non- d4T regimen Estimated weeks (CI)
Southern Africa	59.92 (57.56, 62.27)	133.82 (131.08, 136.56)
East Africa	52.92(50.76, 55.08)	84.88 (80.57, 89.19)
West Africa	43.94 (39.43, 48.45)	92.87 (86.59, 99.14)
Asia-Pacific	38.94 (34.45, 43.43)	54.92 (46.69, 63.15)
Central Africa	61.92 (54.86, 68.98)	60.92 (53.23, 68.37)

Note: Estimates and CI are from PSR model

Similar analyses are presented in Figure 5-3 to Figure 5-6, where the results of the PS-SiZer analysis are presented for the East-Africa, West Africa, Central-Africa, and Asia-Pacific IeDEA regions, respectively. The PS-SiZer maps corresponding to the East and West Africa regions are very similar. For d4T-containing regimens (panel a1 in Figure 5-3 and Figure 5-4), blue areas are followed by purple areas after about 50-60 weeks for most

levels of smoothing, indicating significantly increasing weight during this period. After this point, weight gain diminishes. By contrast, the blue areas in the PS-SiZer maps corresponding to the non-d4T-containing regimens (panels' a2 in Figure 5-3 and Figure 5-4) extend past week 60, indicating that weight continues to increase past 60 weeks after initiation of ART.

Analyses at the optimum smoothing level produced the estimated curves of weight measurements shown in panels' b1 and b2 of Figure 5-3 and Figure 5-4 and in Table 5-2 (rows 2 and 3).

For East Africa, results at the optimal smoothing levels showed that the weight in patients treated with d4T-containing regimens did not significantly increase after 52.9 weeks compared to 84.9 weeks for patients treated with non-d4T-containing regimens. For West Africa, the results are similar, patients treated with d4T-containing regimens estimated to weight gain for 43.9 weeks versus 92.9 weeks for the non-d4T-containing regimens

Analyses of data from the Central Africa IeDEA region are shown in Figure 5-5 (panels b1 and b2) and in Table 5-2 (row 4). The estimated duration of weight increases in the Central Africa region was 61.9 weeks for d4T-containing regimens versus 60.9 weeks for non-d4T-containing regimens. Results from the analyses of data in the Asia Pacific IeDEA region are presented in Figure 5-6 and Table 5-2 (row 5). The estimated duration of weight gain in d4T-containing regimens was 38.9 weeks versus 54.9 weeks in non-d4T-containing regimens.

#### 5.4 Discussion and Conclusions

This study is the largest of its kind ever performed in this context involves data from about 185,010 patients with more than two million clinic visits during the four years of follow-up after initiation of ART treatment. PS-SiZer Maps were presented for Southern Africa, East-Africa, West Africa, Central-Africa, and Asia-Pacific for the IeDEA cohorts. The major finding was that adult HIV-infected patients starting ART with d4T-containing regimens experienced weight gains whose durability was significantly shorter (small blue areas followed by purple areas) than patients who started ART with regimens that did not contain d4T.

The final analysis to detect the duration of significant weight gain at an optimum level shows that weight plateaued after 59.92 weeks (95% CI: 57.56, 62.27) in d4t-treated patients compared to 133.82 weeks (95% CI: 131.08, 136.56) in non-d4T treated patients in Southern Africa region. The difference between the two treatment regimens is significant as the 95% CI did not overlap with each other. The durability of weight gains was significantly shorter for patients treated with regimens containing d4T than patients who started treatment containing a non-d4T regimen for East Africa [(d4T: 52.92 (50.76, 55.08) versus non-d4T: 84.88 (80.57, 89.19)]; West Africa [d4T: 43.94 (39.43, 48.45) versus non-d4T: 92.87 (86.59, 99.14)]; Asia Pacific [d4T: 38.94 (34.45, 43.43) versus non-d4T: 54.92 (46.69, 63.15)]. The weight increase lasted for 61.92 weeks (54.86, 68.98) for d4T-treated vs. 60.92 weeks (53.23, 68.37) for non-d4T treated patients in Central Africa region.

Our analysis showed an unequivocal difference in the pattern of weight changes after ART initiation about whether the first-line regimen included the drug d4T or not. While increases in weight may have been faster in the d4T group, the long-term durability of

weight gain seen in adult HIV-infected patients initiating ART with d4T-containing versus non-d4T containing regimens was much shorter depending on the geographical region. It may be argued that weight is only a partial marker of ART effectiveness, and, further, that focusing on d4T use is not as relevant given recent treatment guidelines. However, d4T is still a mainstay drug in many parts of the world, and weight is a relevant biomarker that is both associated with clinical outcome and is correlated to other biomarkers that are more difficult or expensive to obtain. Regardless, the PS-SiZer methodology presented here is applicable to many other settings and biomarkers collected longitudinally in order to assess the clinical state of patients and the effectiveness of the antiretroviral therapy provided to them.

In conclusion, we detected a relatively shorter durability of weight gain for patients who were treated with d4t-containing regimen consistently among the cohort of patients in all five regions in IeDEA. The change of body weight for HIV-infected patients needs to be monitored closely after initiation of ART treatment, specifically for those containing d4T regimens. After the detection of initial weight loss, a caregiver should consider alternative treatment options. Early detection of weight loss or non-increasing weight after initiation of ART may prevent long-term complication or death for HIV-infected patients. Hence, the overall patient management and assessment of the efficacy of the treatment programs become more effective in resource-limited countries.

## 5.5 Limitations

Along with the important observations presented above, the present study has a number of limitations that need to be considered carefully.

A characteristic of the current analysis is that changes in ART regimens were not taken into consideration. For example, changes from d4T-containing to d4T-non-containing regimens during the follow-up period (dropouts) or vice versa (drop-ins) were not considered. It is unclear whether the former change is more frequent compared to the latter. When the changes from a d4T to a non-d4T-containing regimen are more common, then the induced bias generated by ignoring regimen changes after baseline will tend to attenuate the differences between the two groups. In that case, the difference in weight gains seen between d4T-containing versus non-d4T-containing regimens will be underestimated. The same bias would result from switching to second-line therapy. However, the median time to change a regimen (analysis not shown) was one year and involved an exceedingly small number of patients, so its impact on our analysis is expected to have been minimal.

The effect of high rates of loss to follow-up in the two groups may be considered as another limitation of the study. In particular, if toxicity from d4T-containing regimens leads to patients abandoning care altogether, then their weight will not be measured resulting in an upward bias when estimating the overall weight changes in this group. This bias will have an attenuating effect on the comparison of weight gains between the two groups. However, this does not appear to be the problem in this analysis, as the rates of patient loss to follow-up are not markedly different between the two groups (analysis not shown).

## 5.6 Strengths

Along with these limitations, the study has considerable strengths. The size of the data alone makes this a definitive study for the period ending just prior to the inception of

the WHO 2013 treatment guidelines. Thus, it provides a useful baseline reference of historical data for future studies assessing the effect the changing guidelines had on patient outcomes. In addition, the use of the PS-SiZer methodology provides a useful visual summary of the change over time in important biomarkers, and the smoothing underlying the method provides a strong assurance for the detection of important features in the longitudinal trajectories of these markers. Further, one strength of the study lies in about the extremely large number of participants. Data used in the present study were based on more than 185,000 adult HIV-infected patients recruited from five regions around the world, providing well over two million longitudinal weight measurements.

#### Acknowledgments

We are grateful to all patients, care givers, and data managers involved in the participating cohorts and treatment programs. This work was supported by a contract provided by the Joint United Nations Program on HIV/AIDS (UNAIDS).

Grants: The standardized data collected in the IeDEA consortium was funded by the NIH National Institute of Allergies and Infectious Diseases (NIAID). The work was supported by the following grants: NIH-NIAIDU01AI069911 (EA), NIH-NIAIDU01AI069924 (SA), NIH-NIAIDU01AI069919 (WA), and NIH-NIAID U01AI069907 (AP).

Research Grant: Samiha Sarwat was supported by NIH Grant Number TL1 TR000162 (A. Shekhar, PI) during July, 2012 to June, 2014.

Appendix 5A: PS-SiZer maps for East, West and Central Africa, and Asia Pacific.

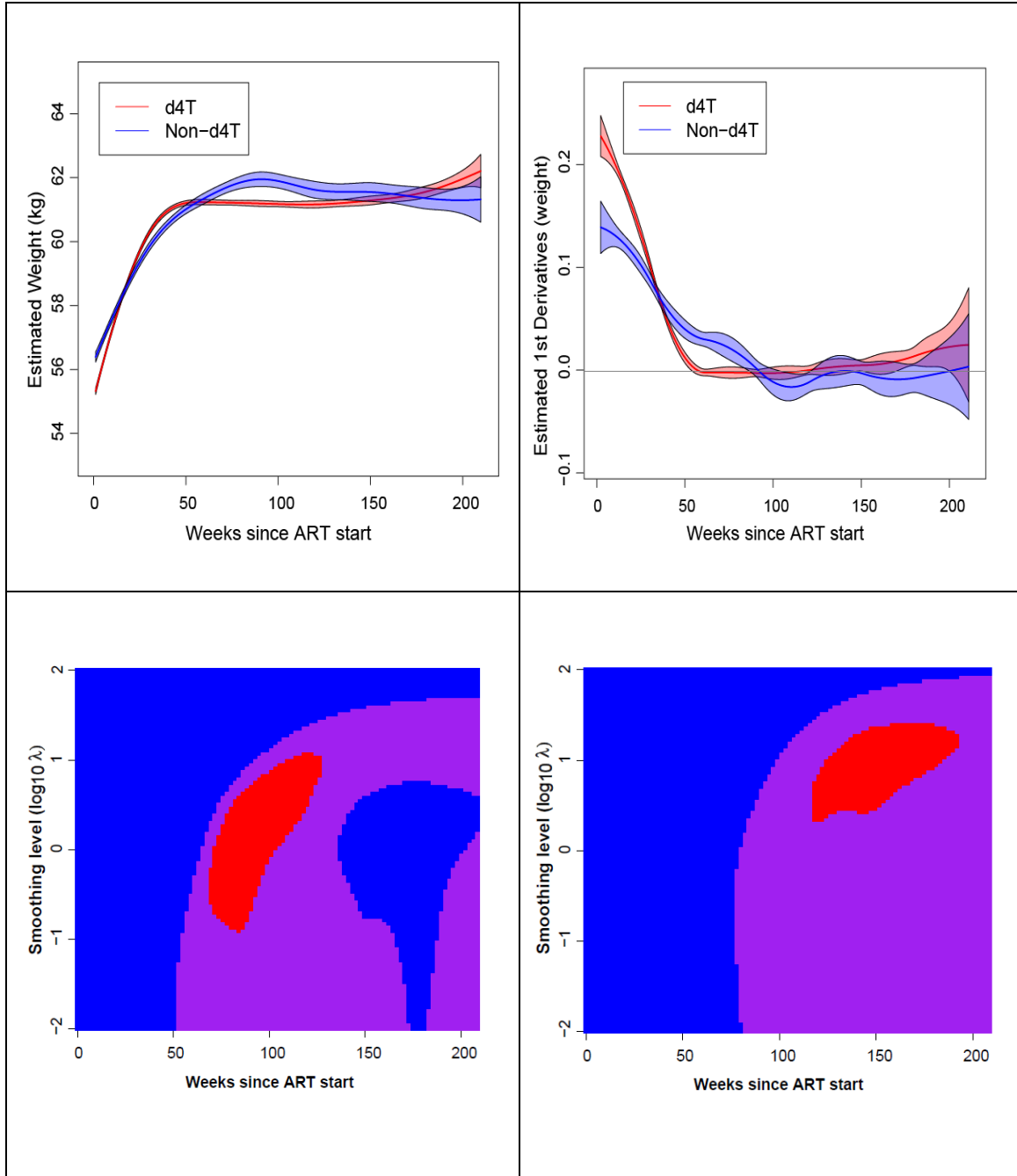


Figure 5-3 Top-row: left panel represents estimated the weight change with CL, and the right panel represents the estimated first derivatives with CL over time by d4T and non-d4T regimens at the REML optimum level. Bottom- row: left panel represents the PS-SiZer map for d4T-regimen, and the right panel represents the PS-SiZer map for non-d4T-regimen. The vertical axis represents the level of smoothing expressed in  $10 \lambda$ , and the horizontal axis represents the time in weeks since the start of ART. Data is from East Africa region.



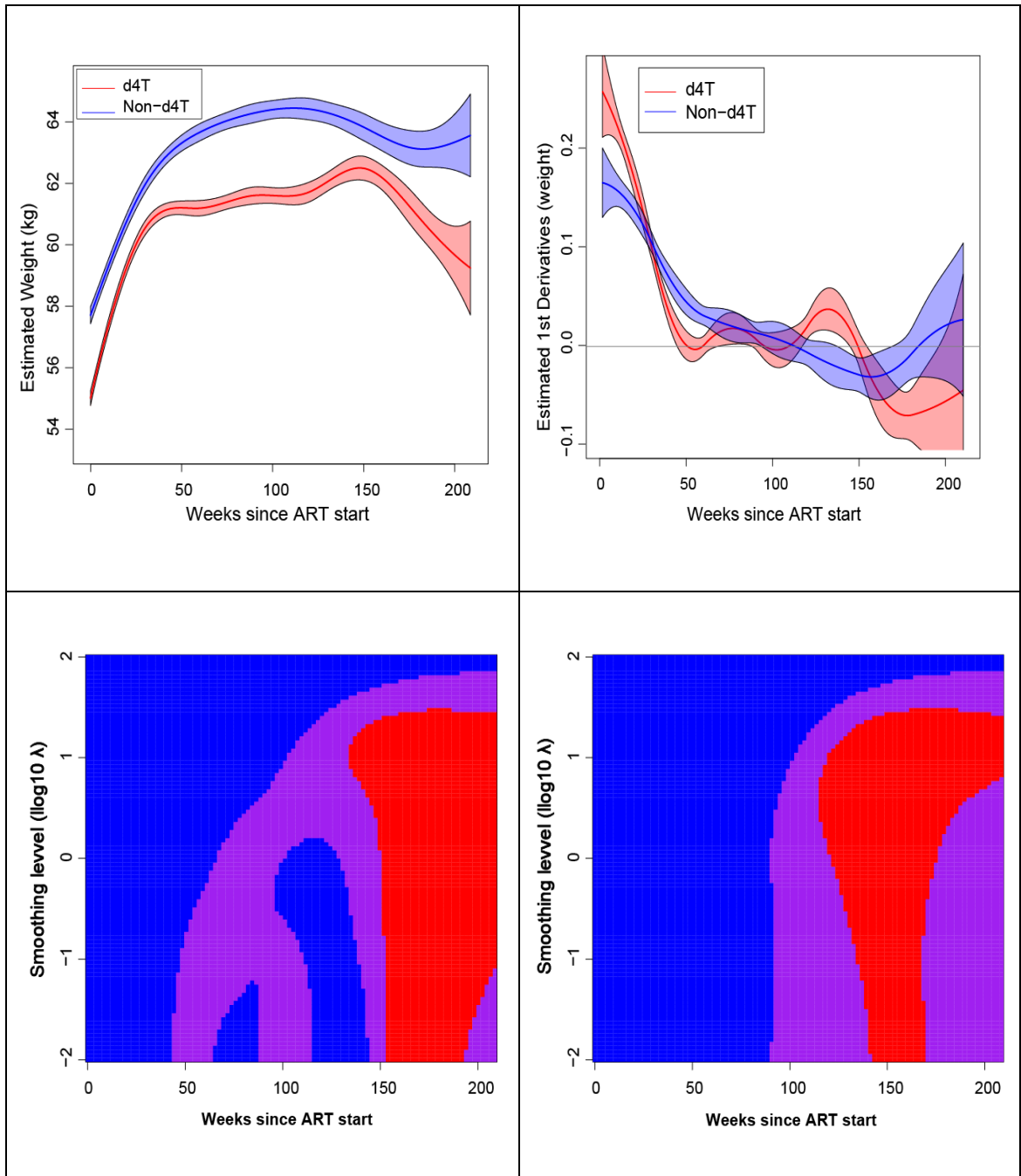


Figure 5-4 Top-row: left panel represents the estimated weight change with CL, and the right panel represents the estimated first derivatives with CL over time by d4T and non-d4T regimens at the REML optimum level. Bottom-row: left panel represents the PS-SiZer map for d4T-regimen, and the right panel represents the PS-SiZer map for non-d4T-regimen. The vertical axis represents the level of smoothing expressed in  $10 \lambda$ , and the horizontal axis represents the time in weeks since the start of ART. Data is from West Africa.

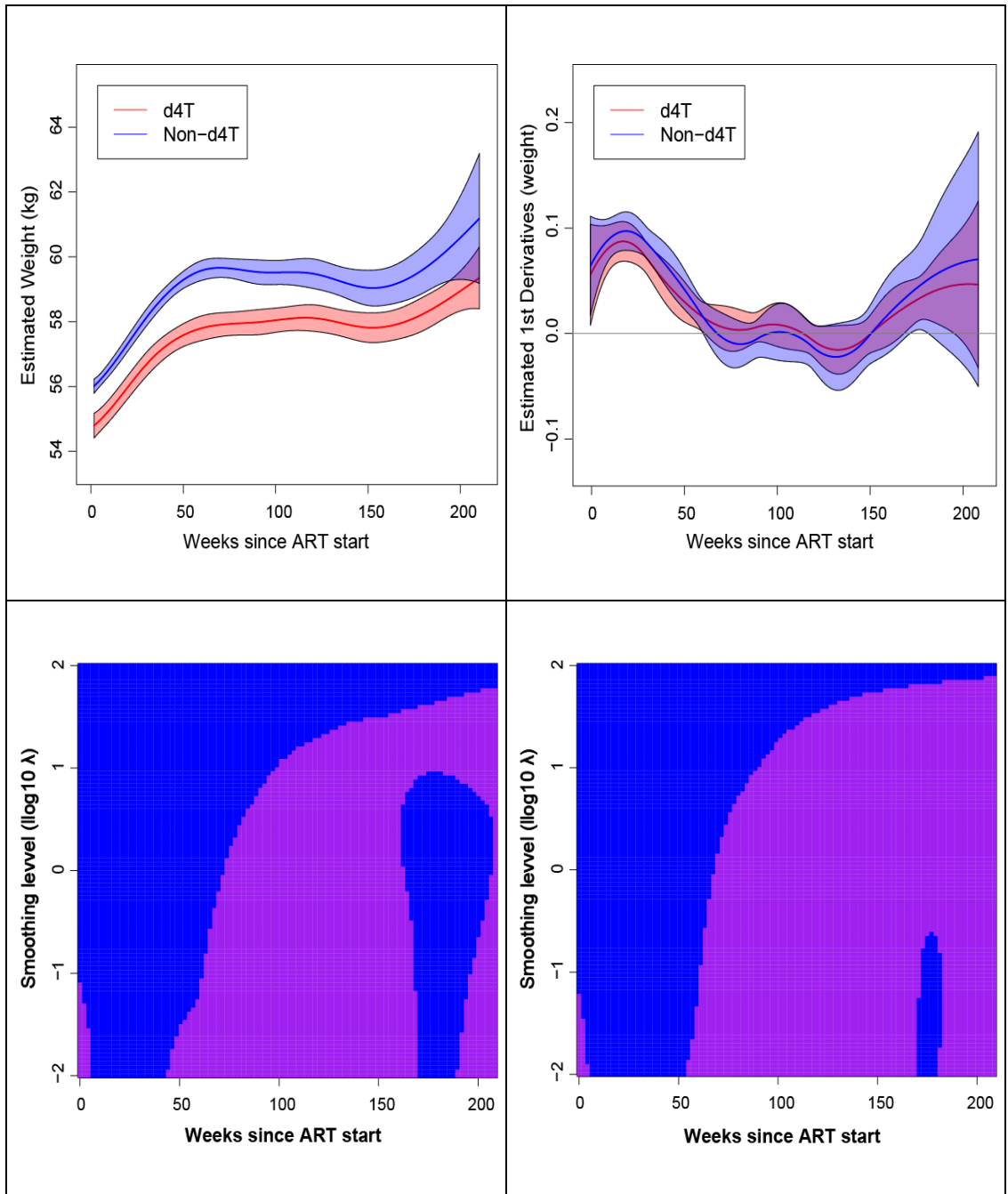


Figure 5-5 Top-row: left panel represents the estimated weight change with CL, and the right panel represents the estimated first derivatives with CL over time by d4t and non-d4T regimens at the REML optimum level. Bottom- row: left panel represents PS-SiZer map for d4T-regimen, and the right panel represents the PS-SiZer map for non-d4T-regimen. The vertical axis represents the level of smoothing, expressed in  $10 \lambda$ , and the horizontal axis represents the time in weeks since the start of ART. Data is from Central Africa region.

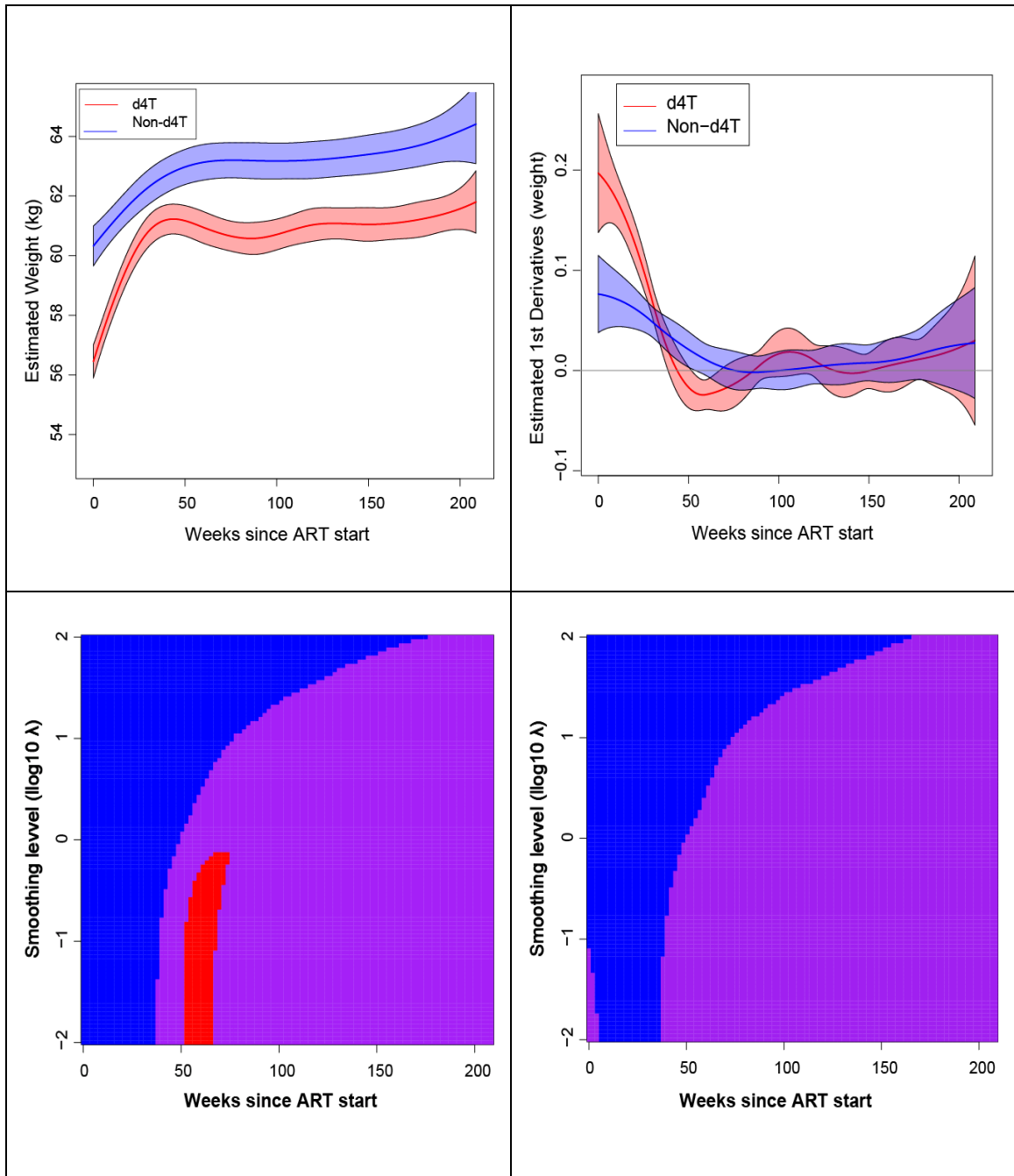


Figure 5-6 Top-row: left panel represents the estimated weight change with CL, and the right panel represents the estimated first derivatives with CL over time by d4T and non-d4T regimens at the REML optimum level. Bottom- row: left panel represents the PS-SiZer map for d4T-regimen, and the right panel represents the PS-SiZer map for non-d4T-regimen. The vertical axis represents the level of smoothing expressed in  $10 \lambda$ , and the horizontal axis represents the time in weeks since the start of ART. Data is from Asia Pacific region.

## REFERENCES

- Biadgilign, S., Reda, A., & Digafe, T. (2012). Predictors of mortality among HIV infected patients taking antiretroviral treatment in Ethiopia: a retrospective cohort study. *AIDS Res Ther*, 9:15.
- Boulle, A., Schomaker, M., May, M. T., Hogg, R. S., Shepherd, B., Monge, S., & Sterne, J. A. (2014). Mortality in Patients with HIV-1 Infection Starting Antiretroviral Therapy in South Africa, Europe, or North America: A Collaborative Analysis of Prospective Studies. *PLoS Med*, 11(9): e1001.
- Brumback, B., Ruppert, D., & Wand, M. (1999). Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior: Comment. *Journal of the American Statistical Association*, 94:794–797.
- Chaudhuri, P., & Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447), 807-823.
- Egger, M., Ekouevi, D., Williams, C., Lyamuya, R., Mukumbi, H., Braitstein, P., Wools-Kaloustian, K. (2012). Cohort Profile: The international epidemiological databases to evaluate AIDS (IeDEA) in sub-Saharan Africa. *Int J Epidemiol*, 41:1256–64.
- Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–102.
- Gallant, J. E., Staszewski, S., Pozniak, A. L., DeJesus, E., Suleiman, J. M., Miller, M. D., & Group (2004). Efficacy and safety of tenofovir DF vs stavudine in combination therapy in antiretroviral-naïve patients: a 3-year randomised trial. *JAMA*, 292: 191-201.

- Griensven, v. J., De Naeyer, L., Mushi, T., Ubarijoro, S., Gashumba, D., Gazille, C., & Zachariah, R. (2007). High prevalence of lipoatrophy among patients on stavudine-containing first-line antiretroviral therapy regimens in Rwanda. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 101(8), 793-798.
- Griensven, v. J., Zachariah, R., Mugabo, J., & Reid, T. (2010). Weight loss after the first year of stavudine containing antiretroviral therapy and its association with lipoatrophy, virological failure, adherence and CD4 counts at primary health care level in Kigali, Rwanda. *Trans R Soc Trop Med Hyg*, 104(12):751-7.
- Grinspoon, S., & Mulligan, K. (2003). Weight loss and wasting in patients infected with human immuno deficiency virus. *Clin Infect Dis*, 36: S69–78.
- IeDEA. (n.d.). *IeDEA*. Retrieved Jan 2015, from International Epidemiologic Databases to Evaluate AIDS: <http://www.iedea.org/>
- Jaroslaw, H., Elena, N., & Nan, L. M. (2007). Long CriSP: A test for bump hunting in longitudinal data. *Statistics in Medicine*, 26: 1383-1397.
- Joly, V., Flandre, P., Meiffredy, V., Leturque, N., Harel, M., Aboulker, J., & Yeni, P. (2002). Increased risk of lipoatrophy under stavudine in HIV-1-infected patients: results of a substudy from a comparative trial. *AIDS*, 16(18):2447-54.
- Li, N., Spiegelman, D., Drain, P., Mwiru, R. S., Mugusi, F., Chalamilla, G., & Fawzi, W. W. (2012). Predictors of weight loss after HAART initiation among HIV-infected adults in Tanzania. *AIDS*, 26(5):577-85.
- Madec, Y., Szumilin, E., Genevier, C., Ferradini, L., Balkan, S., Pujades, M., & Fontanet, A. (2009). Weight gain at 3 months of antiretroviral therapy is strongly associated

- with survival: evidence from two developing countries. *Volume 23* (Issue 7- p 853–861).
- May, M., Boulle, A., Phiri, S., Messou, E., Myer, L., Wood, R., & Egger, M. (2010). Prognosis of patients with HIV-1 infection starting antiretroviral therapy in sub-Saharan Africa: a collaborative analysis of scale-up programmes. *Lancet*, 376: 449–57.
- R Development Core Team. (2008). R: A language and environment for statistical computing. Retrieved from R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0: <http://www.R-project.org>.
- Rosen, S., Long, L., Fox, M., & Sanne, I. (2008). Cost and cost-effectiveness of switching from stavudine to tenofovir in first-line antiretroviral regimens in South Africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 48(3), 334-344.
- Su, X., Simmons, Z., Mitchell, R., Kong, L., Stephens, H., & Connor, J. (2013). Biomarker-based predictive models for prognosis in amyotrophic lateral sclerosis. *JAMA Neurology*, 70(12):1505-11.
- Subbaraman, R., Chaguturu, S., Mayer, K., Flanigan, T., & Kumarasamy, N. (2007). Adverse effects of highly active antiretroviral therapy in developing countries. *Clin Infect Dis*, 45(8):1093-101.
- Tang, A., Jacobson, D., Spiegelman, D., Knox, T., & Wanke, C. (2005). Increasing risk of 5% or greater unintentional weight loss in a cohort of HIV-infected patients, 1995 to 2003. *J Acquir Immune Defic Syndr*, 40(1):70-6.

- UNAIDS. (2013). *Treatment 2015*. Retrieved from UNAIDS Corporate publications :  
[http://www.unaids.org/sites/default/files/media\\_asset/JC2484\\_treatment-2015\\_en\\_1.pdf](http://www.unaids.org/sites/default/files/media_asset/JC2484_treatment-2015_en_1.pdf)
- WHO. (2003). *Working Document on Monitoring and Evaluating of National ART Programmes in the Rapid Scaleup to 3 by 5*. Retrieved from The 3 by 5 initiative:  
<http://www.who.int/3by5/publications/documents/artindicators/en/index.html>
- WHO. (2009). *Towards universal access: scaling up priority HIV/AIDS interventions in the health sector*. Geneva.
- WHO. (2013). *Global update on HIV treatment: results, impact and opportunities*. Geneva.  
Retrieved from World Health Organisation.
- Wood, S. (2010). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*. Retrieved from <http://cran.r-project.org/package=mgcv>.
- Wools-Kaloustian, K., Kimaiyo, S., Diero, L., Siika, A., Sidle, J., Yiannoutsos, C. T., & Tierney, W. M. (2006). Viability and effectiveness of large-scale HIV treatment initiatives in sub-Saharan Africa: experience from western Kenya. *AIDS*, 20:41–48.

## CURRICULUM VITAE

**Samiha Sarwat**

### **Education**

**PhD in Biostatistics** (2015)

Indiana University, Indiana University Purdue University Indianapolis, IN

**PhD Minor:** Public Health

**Master of Science in Applied Statistics** (May, 1999)

Department of Mathematics & Statistics, University of Arkansas, Fayetteville

**Bachelor of Science in Statistics** (1996)

Department of Statistics, University of Dhaka, Dhaka, Bangladesh

**Fellowship** (Jun 2012 - Jun 2014): Clinical and Translational Sciences Institute (CTSI)

pre-doctoral Trainee (TL1 Program: A. Shekhar, PI).

**PhD Research:** Penalized Spline modeling of the ex-vivo assays dose-response curves and the HIV-infected patients' bodyweight change.

### **Professional Experience**

**Jun 2014 – Present**                      **Senior Biostatistician**

**Quintiles, North Carolina, USA**

Primarily responsible for developing protocols, preparing statistical analysis plans, and writing statistical section of the clinical study report.

**Oct' 2010 to 2012**                      **Senior Biostatistician**

**i3Statprobe, Indianapolis, IN, USA**

Served as a diabetes therapeutic expert and provided statistical guidance to ensure deliverables are consistent, accurate, and adhere to the clinical strategy.



**Mar 2003 to Oct 2010**      **Project Statistician**

**Eli Lilly and Company, Indianapolis, IN, USA**

Served as a senior level statistician in Diabetes and Insulin-Device clinical trials.

Worked as a project statistician to support clinical trials activities, sample size calculation, writing statistical analysis plan and running the statistical analysis in the Acute Care (Infectious disease) product team.

**Oct 2001-Oct 2003**      **Biostatistician**

**Medfocus, Indianapolis, Indiana, USA**

Worked as a statistical programmer and analyst

Co-authored several published manuscripts and participated in the analyses that lead to several additional publications.

**Jul 1999-Dec 2000**      **Statistician**

**University of Arkansas Medical School (UAMS),**

**Center for outcome research and Outcomes,**

**Little Rock, Arkansas, USA**

Worked as a primary statistician in studies related to outcome research including Depression and Schizophrenia clinical trials, health outcome data and Medicaid database

**Jan 98-May 99**      **Teaching Assistant**

**Department of Mathematics & Statistics,**

**University of Arkansas, Fayetteville, Arkansas**

**Professional Affiliation**      American Statistical Association, USA

**Technical Skills**      General skills in statistics and statistical computing

## Publications

### Oral Poster Presentation

EASD, *Denmark, 2006*. “Hyperglycemia and its Effect After Acute Myocardial Infarction on Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus (HEART2D): Regional Differences in Baseline CV Risk Factors and Treatments.”

EASD, *Sweden, 2010* ‘Comparison of Insulin Diluent Leakage Post Injection Using Two Different Needle Lengths and Injection Volumes in Obese Patients with Type 1 or Type 2 Diabetes Mellitus’.

### Manuscripts

1. Jean-Louis Vincent, et.al, **Samaha Sarwat**, MS. Drotrecogin alfa (activated) in patients with severe sepsis presenting with purpura fulminans, meningitis, or meningococcal disease: a retrospective analysis of patients enrolled in recent clinical studies. *Critical Care* 2005.
2. Didier Payen, MD, PhD, Armin Sablotzki, MD, PhD, Philip S. Barie, MD, MBA, Graham Ramsay, MD, PhD, Stephen Lowry, MD, Mark Williams, MD, **Samaha Sarwat**, MS, Justin Northrup, MPT, International integrated database for the evaluation of severe sepsis and drotrecogin alfa (activated) therapy: Analysis of efficacy and safety data in a large surgical cohort. *Surgery* 2006.
3. Daniel J. Cox PhD, Anthony McCall, MD, PhD, Boris Kovatchev PhD, **Samaha Sarwat** MS, Liza L. Ilag MD, Meng H. Tan MD Effects of Blood Glucose Changes on Perceived Mood and Cognitive Symptoms in Insulin-treated Type 2 Diabetes. *The Diabetes Care* 2007.

4. Zvonko Milicevic, Itamar Raz, Scott D Beattie, Barbara N Campaigne, Samiha Sarwat, Elwira Gromniak, Irina Kowalska, Edvard Galic, Meng Tan, Markolf Hanefeld. Natural history of cardiovascular disease in patients with diabetes: Role of hyperglycemia. *Diabetes Care (supplement) 2007*.
5. David C. Robbins; Paul J. Beisswenger, Antonio Ceriello, Ronald B. Goldberg, Robert G. Moses, Emmanuil M. Pagkalos, Zvonko Milicevic, Cate A. Jones, **Samiha Sarwat**, Meng H. Tan.. Achievement of HbA1c, Pre-prandial and Postprandial Blood Glucose Targets - A 24-week, Parallel-group, Multi-country, Open-labeled, Randomized Comparison of a Mealtime 50:50 Basal+Prandial Insulin Analog Mixture Regimen with a Basal Insulin Analog Regimen, both Plus Metformin in Type 2 Diabetes Patients. *Clinical Therapeutics, 2007*.
6. LevineRL, LeClerc JR, Bailey JE, Monberg MJ, Sarwat S. Venous and arterial thromboembolism in severe sepsis. *Thrombosis and Haemostasis, 2008*.
7. R. Beale, K. Reinhart, F. Brunkhorst, G. Dobb, M. Levy, G. Martin, C. Martin, G. Ramsey, E. Silva, B. Vallet, J.L. Vincent, J.M. Janes, S. Sarwat, M.D. Williams. PROGRESS (Promoting Global Research Excellence in Severe Sepsis): Lessons from an international sepsis registry, *Infection 2008*.
8. Debra A. Ignaut, Sherwyn L. Schwartz, **Samiha Sarwat**, Heather L. Murphy. Comparative Device Assessments: Humalog KwikPen Compared with Vial & Syringe and FlexPen" *Diabetes Educator (Sept 2009)*.
9. David Shrom, **Samiha Sarwat**, Liza Ilag and Zachary T. Bloomgarden. Does A1c consistently reflect mean plasma glucose?. *Journal of Diabetes, 2010*.

10. Antonio Roberto Chacra, Mark Kipnes, Liza L. Ilag, **Samiha Sarwat**, Joseph Giaconia and John Chan. Comparison of Insulin Lispro Protamine Suspension and Insulin Detemir in Basal-Bolus Therapy in Patients with Type 1 Diabetes. *Diabetic Medicine*, 2010.
11. Paula E. Clark, Virginia Valentine, Jennifer N. Bodie, **Samiha Sarwat**. Ease of use and patient preference pens injection simulation study comparing two prefilled insulin. *Current Medical Research & Opinion*, 2010,
12. **S. Sarwat**, L. L. Ilag, M. A. Carey, D. S. Shrom, R. J. Heine. The Relationship between Self-Monitored Blood Glucose Values and Glycated haemoglobin in Insulin-treated Patients with Type 2 Diabetes. *Diabet Med* 2010;27.
13. P.J. Beisswenger, W.V. Brown<sup>2</sup>, A. Ceriello, N.A. Le<sup>2</sup>, R.B. Goldberg, J.P. Cooke, D.C. Robbins, **S. Sarwat**, H. Yuan, C.A. Jones, M.H. Tan. Meal-Induced Increases in C-Reactive Protein, Interleukin-6, and Tumor Necrosis Factor  $\alpha$  are Attenuated by Prandial+Basal Insulin in Type 2 Diabetes Patients. *Diabet Med* 2011.

### **Presentations & Abstracts**

1. S. Betty Yan, Bruce Basson, John Brandt, Jean-Francois Dhainaut, David Joyce, **Samiha Sarwat**. Survival advantage associated with Factor V Leiden mutation in patients with severe sepsis in the PROWESS study. *Gordon Conference*, 2002.
2. Robert Levine, Jacques Leclerc, Joan Bailey, Matthew Monberg, **Samiha Sarwat**. Incidence Rates of Venous and Arterial Thromboembolism in Severe Sepsis. *ACCP Oct 2003*.

3. Payen D, Williams M D, Sarwat S, Janes J. Safety and Efficacy of Drotrecogin Alfa (Activated) in Adult Surgical Patients with Severe Sepsis. *ESICM 2004*.
4. Rekha Garg, Simon Nadel, Sau Chi Betty Yan, Virginia Wyss, **Samiha Sarwat**, Carol Mitchell, Joan Bailey, Stephen Shinall, Jonathan Janes. Pediatric severe sepsis patients with or without purpura fulminans had similar outcomes. *ICCAC 2004*.
5. Richard Beale, Konrad Reinhart, Rekha Garg, Eliezer Silva, Geoffrey Dobb, **Samiha Sarwat**, Jean-Louis Vincent. PROGRESS Severe Sepsis Registry Data Indicates Mortality from Severe Sepsis Remains high. *ESICM 2004*.
6. Rekha Garg, Richard Beale, Eliezer Silva, Konrad Reinhart, Geoffrey Dobb, **Samiha Sarwat**, and Jean-Louis Vincent ICU Mortality in Severe Sepsis Patients Varies in Countries Participating in PROGRESS. *ESICM 2004*.
7. Richard Beale, Konrad Reinhart, Eliezer Silva, Geoffrey Dobb, **Samiha Sarwat**, Rekha Garg, Jean-Louis Vincent. Comparison of PROGRESS Severe Sepsis Registry Patients to INDEPTH Integrated Clinical Trial Database Placebo Patients. *CHEST 2004*.
8. Konrad Reinhart, Frank Brunkhorst, Eliezer Silva, Geoffrey Dobb, Jean-Louis Vincent, **Samiha Sarwat**, Rekha Garg, Richard Beale. Geographic Variations in Use of Recommended Severe Sepsis Interventions Observed in PROGRESS Severe Sepsis Registry Data. *CHEST 2004*.
9. Scott J Jacober, Paul J Beisswenger, Robert G Moses, Cate A Jones **Samiha Sarwat** David C Robbins, Meng H Tan. Lispro Mid Mixture (MM) Plus Metformin (Met) Reduced Mean Daily and Pre-/Post- Meal Blood Glucose (BG)

and Their Excursions (Exc) More Than Glargine (G) Plus Met in Type 2 Diabetes (T2D) Patients (Pts). *ADA 2006*.

10. Meng H Tan, W Virgil Brown, Ronald B Goldberg, Ngoc-Anh Le, Antonio Ceriello, Paul J Beisswenger, Cate A Jones, **Samiha Sarwat**, David C Robbins. Lispro Mid Mixture (MM) Plus Metformin (Met) Reduced Postprandial hs-CRP More Than Glargine (G) Plus Met in Type 2 Diabetes (T2D) Patients (Pts) . *ADA 2006*.
11. David C Robbins, Paul J Beisswenger, Antonio Ceriello, **Samiha Sarwat**, Cate A Jones, Meng H Tan. Thrice-daily Lispro Mid Mixture (MM) Plus Metformin (Met) Improved Glycemic Control Better Than Glargine (G) Plus Met in Type 2 Diabetes (T2D) Patients (Pts). *ADA 2006*.
12. Zvonko Milicevic, David Oakley, **Samiha Sarwat**, Meng H. Tan, David C. Robbins. Hyperglycemia and its Effect after Acute myocardial infarction on cardiovascular outcomes in patients with Type 2 Diabetes Mellitus (HEART2D): Patient Population Baseline Characteristics. *ADA 2006*.
13. Zvonko Milicevic, Barbara N. Campaigne, **Samiha Sarwat**, Meng H. Tan, David C. Robbins. Hyperglycemia and its Effect after Acute myocardial infarction on cardiovascular outcomes in patients with Type 2 Diabetes Mellitus (HEART2D): Baseline Cardiovascular Treatments and Risk Factors
14. David C Robbins, Paul J Beisswenger, Robert G Moses, Antonio Ceriello, Zvonko Milicevic, **Samiha Sarwat**, Cate A Jones, Meng H Tan .Comparison of Insulin Lispro Mid Mixture (MM) Plus Metformin (Met) with Glargine (G) Plus

Met on HbA1C and Blood Glucose (BG) Profiles in Patients With Type 2 Diabetes (T2D). *EASD 2006*.

15. Meng H Tan<sup>1</sup>, Ngoc-Anh Le<sup>2</sup>, Ronald B Goldberg<sup>3</sup>, W Virgil Brown<sup>2</sup>, Antonio Ceriello<sup>4</sup>, Paul J Beisswenger<sup>5</sup>, **Samiha Sarwat**<sup>1</sup>, Cate A Jones<sup>1</sup>, Zvonko Milicevic<sup>1</sup>, David C Robbins. Postprandial Increase in hs-CRP in Patients with Type 2 Diabetes (T2D): Comparison of Lispro Mid Mixture (MM) Plus Metformin (Met) with Glargine (G) plus Met. *EASD 2006*.
16. Zvonko Milicevic, Itamar Raz, **Samiha Sarwat**, Velimir Bozиков, Puvanesveri Naiker, Krzysztof Strojek, David Oakley, David C. Robbins, Meng H. Tan. Hyperglycemia and its Effect After Acute Myocardial Infarction on Cardiovascular (CV) Outcomes in Patients with Type2 Diabetes Mellitus (HEART2D): Baseline Characteristics. *EASD 2006*.
17. Zvonko Milicevic, Barbara Campaigne, **Samiha Sarwat**, David C. Robbins, Elwira Gromniak, Irina Kowalska, Izet Aganovic, Meng Tan. Hyperglycemia and its Effect After Acute Myocardial Infarction on Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus (HEART2D): Regional Differences in Baseline CV Risk Factors and Treatments. *EASD 2006*.
18. Meng H Tan, Ngoc-Anh Le, Ronald B Goldberg, Antonio Ceriello, Paul J Beisswenger, **Samiha Sarwat**, David C Robbins, W Virgil Brown. Comparison of lispro mid mixture (MM) plus metformin (Met) with glargine (G) plus Met on postprandial increase in hs-CRP in patients with type 2 diabetes (T2D). *IDF 2006*.

19. Daniel J. Cox, Anthony McCall, Boris Kovatchev, Liza L. Ilag, **Samiha Sarwat**, Meng H. Tan. Mood and Cognitive Effects of Insulin Treatment (Tx) in Type 2 Diabetes (T2D) Patients (Pts). *ADA2007*.
20. Meng H Tan, W Virgil Brown, Ronald B Goldberg, Antonio Ceriello, Paul J Beisswenger, Ngoc-Anh Le, **Samiha Sarwat**, Cate A Jones, Zvonko Milicevic, David C Robbins. Postprandial increase in IL 6 and TNF $\alpha$  is reduced by a basal+prandial insulin regimen compared with a basal insulin regimen in type 2 diabetes (T2D) patients. *ADA 2007*.